EHR-TASR: Leveraging Time-Aware Scoring and LLM Reasoning for Clinical Prediction

Xiaoyang Meng[1,3], Xiaoding Zhou[1,3,✉], Yang Liu[1,3], Weiyu Zhang[1,3,✉], Hongjiao Guan[1,3], Xueping Peng[2], and Wenpeng Lu[1,3]

[1]*Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China*

[2]*Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia*

[3]*Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China*

✉ *Corresponding author: 15953123827@126.com, zwy@qlu.edu.cn*

***Abstract*—Clinical prediction tasks are critical for supporting medical decision-making in healthcare. Recent advances in large language models (LLMs) have opened new opportunities to improve predictive accuracy and clinical reasoning. However, existing methods are limited in capturing both short-term fluctuations and long-term trends in physiological signals, and insufficient interpretability remains a key barrier to clinical adoption. To address these challenges, we propose EHR-TASR, a novel framework that integrates temporal dynamics with interpretable reasoning for clinical prediction. Specifically, we segment patient visits into 12-hour windows and compute multi-scale time-aware scores to capture deviations in vital signs, thereby supporting the modeling of physiological changes across different time scales. These temporal features are combined with structured electronic health record (EHRs) data and processed by a Bi-LSTM with self-attention to enhance sequential representation. In parallel, a locally fine-tuned LLM generates structured reasoning chains from medical concepts, ensuring interpretability and privacy preservation. Finally, the outputs of the sequence model and the LLM are integrated through a linear combination to generate robust predictions. Experimental results on the public MIMIC-III and MIMIC-IV datasets demonstrate that EHR-TASR consistently outperforms state-of-the-art baselines.**

***Index Terms*—healthcare, explainable AI, clinical prediction, electronic health records.**

I. Introduction

Electronic Health Records (EHRs) store comprehensive and detailed data about patients' diagnoses and treatments, including basic information, medical history, diagnoses, laboratory results, and vital signs, thus providing abundant clinical cues for machine-learning models [1], [2]. However, compared with "dense" data such as images or speech, EHRs are inherently sparse and irregular. The visit frequencies differ markedly among patients, and even for a single patient the timestamps of tests, procedures, and orders are highly uneven, resulting in variable sequence lengths and numerous missing values. These multi-source, heterogeneous, and time-sparse characteristics greatly increase modeling difficulty and limit the model's generalization and interpretability [3].

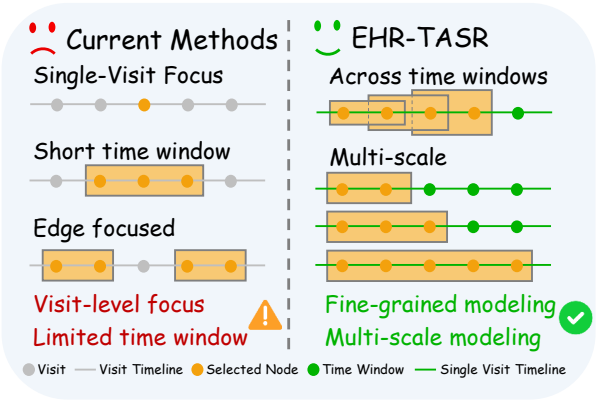

Fig. 1: Comparison between the time-aware scoring method and prior approaches.

In recent years, modeling longitudinal EHRs has progressed rapidly, leading to substantial performance improvements in downstream clinical prediction tasks [4]. A prominent research line focuses on prototypical representation learning to enhance robustness under data sparsity with representative methods including PPN [5] and PRISM [6]. In parallel, emerging frameworks like RAM-EHR [7] and EMERGE [8] utilize the reasoning capabilities of large language models (LLMs) to enrich clinical representations with deeper medical semantics.

However, existing methods still face two key challenges in real-world healthcare applications [9]. **Challenge 1:** Models lack the ability to capture fine-grained fluctuations of patients' physiological indices across time periods. Prior approaches typically treat a single visit as the minimal time unit, focus only on the earliest, most recent, or several intermediate visits, and supplement this with positional or relative-time encodings, as illustrated in Fig. 1. Such practice fails to fully portray the dynamic changes and cumulative effects of continuous

This work is supported by the Program of New Twenty Policies for Universities of Jinan (No.202333008), Program of Innovation Improvement for Small and Medium-sized Enterprises of Shandong (No.2023TSGC0274), the Pilot Project for Integrated Innovation of Science, Education, Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2025ZDZX01), and Shandong Talent Introduction Program (No.WSR2025005).

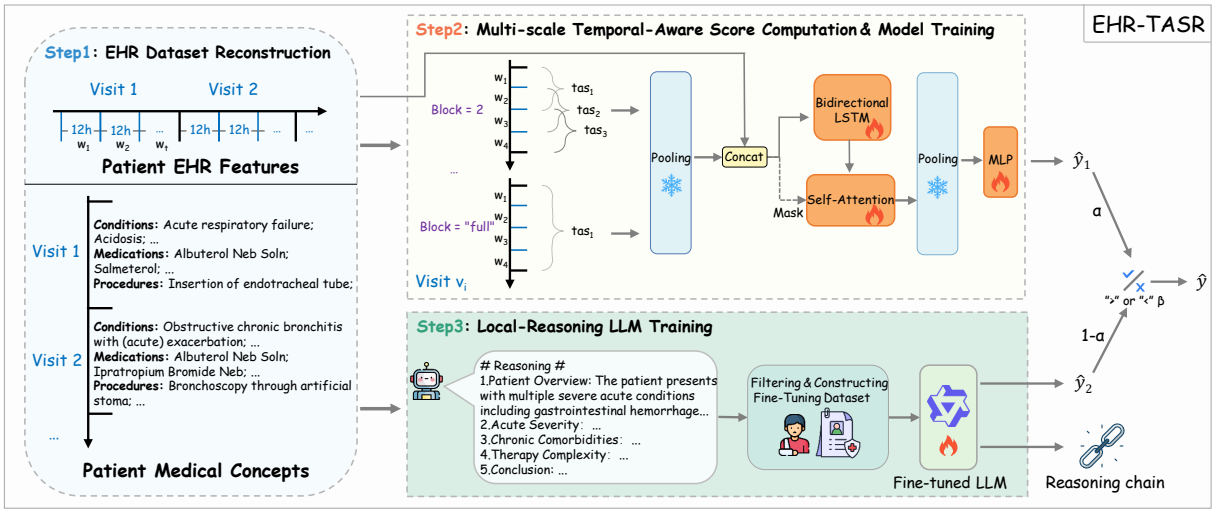

Fig. 2: Overall architecture of our proposed EHR-TASR framework comprises three stages: (1) EHR Dataset Reconstruction, (2) Multi-scale Temporal-Aware Score Computation & Model Training, and (3) Local-Reasoning LLM Training.

measures, such as glucose and blood pressure, across multiple spatiotemporal scales in chronic disease management and acute exacerbation warning. **Challenge 2:** Mainstream LLMs operate largely as end-to-end black boxes, featuring insufficient interpretability and a risk of data leakage. Medical decision-making relies on interpretable outputs to support physicians' clinical accountability, yet existing models usually provide only probabilities or brief summaries, making it difficult to reveal which course-of-illness information underlies their conclusions, thereby weakening physicians' trust and hindering deployment in high-risk scenarios.

To address the aforementioned challenges, we propose a novel framework, EHR-TASR [1] . As illustrated in Fig. 2, this framework computes multi-scale time-aware scores to augment longitudinal EHR data and integrates a sequence model and locally deployed LLMs to capture fine-grained physiological dynamics while enhancing interpretability through task-aligned reasoning chains. Specifically, for Challenge 1, we segment each patient's visits into finer 12-hour windows, compute time-aware scores across multiple inter-window scales for each visit, and feed these scores together with longitudinal EHR data into a temporal model to strengthen its awareness of physiological variability. For Challenge 2, we locally fine-tune and deploy lightweight LLMs to mitigate data leakage and deployment overhead, while experiments show that these models still achieve strong performance. Meanwhile, through fine-tuning and prompts, we guide the model to output physician-friendly reasoning chains oriented toward the prediction results. The chains are structured into five components: Patient Overview, Acute Severity, Chronic Comorbidities, Therapy Complexity, and Conclusion. These chains comprehensively summarize the links between historical illnesses and the prediction outcome, strengthening physicians' understanding and trust in the results. In summary, our key contributions are as follows:

- We introduce EHR-TASR, a novel framework that augments longitudinal EHR data by computing multi-scale time-aware scores to capture feature-value changes across finer temporal intervals, thereby encouraging the model to focus on subtle physiological fluctuations during training.
- We integrate the predictions from a sequential model and a locally deployed LLM, and use a predefined prompt to guide the LLM in producing structured, clinician-friendly reasoning chains to enhance clinical decision support and interpretability.
- Extensive experiments conducted on two real-world EHR datasets demonstrate that our method outperforms state-of-the-art models.

II. Related Works

A. Healthcare Prediction with Longitudinal EHR Modeling

Recent advances in healthcare prediction have increasingly focused on modeling longitudinal EHRs to capture temporal dynamics and patient state evolution. DeepMPM [10] demonstrated that attention-based deep models can effectively leverage longitudinal ICU records to improve mortality risk prediction while offering enhanced interpretability. TA-RNN [11] further highlighted the importance of modeling irregular time intervals by introducing time-aware recurrent architectures with diagnostic attention.

More recent studies have explored Transformer-based architectures for longitudinal EHR modeling. Hi-BEHRT [12] employed a hierarchical Transformer to capture event-level and visit-level dependencies in multimodal EHRs, while TransformEHR [13] adopted an encoder–decoder framework to learn latent data distributions for improved disease outcome prediction. To enhance robustness under data sparsity, IGNITE [14] introduced diffusion-based imputation for individualized

[1]https://github.com/Tiome-tt/EHR-TASR-main

missing value modeling, PAI [15] replaced explicit imputation with learnable prompt-based pseudo-inputs, and PRISM [6] leveraged prototype-driven patient representations with missing-feature-aware calibration to improve generalization.

Despite these advances, existing approaches predominantly rely on implicit sequence representations or coarse-grained temporal modeling, which may limit their ability to capture fine-grained physiological variations across short and irregular time intervals.

B. Healthcare Prediction with Large Language Models

With the rapid development of LLMs, recent studies have explored leveraging LLMs in healthcare prediction to enhance reasoning ability and semantic understanding. Kwon et al. [16] proposed a reasoning-aware diagnostic framework that simulates clinical reasoning through prompt generation and retrieval, demonstrating improved diagnostic performance. Zhu et al. [17] showed that carefully designed prompting strategies enable LLMs to perform clinical prediction from structured longitudinal EHR data, highlighting their potential for medical reasoning.

Researchers have further explored retrieval-augmented and knowledge-enhanced frameworks for healthcare prediction. EMERGE [8] enriched EHR representations by retrieving relevant patient history for patient summarization, while RAM-EHR [7] integrated external medical knowledge to bridge structured EHRs with free-text information. In parallel, GraphCare [18] incorporated biomedical knowledge graphs to enhance clinical reasoning and predictive interpretability.

Although these methods improve the accuracy and transparency of clinical prediction models, their explanations are often designed for researchers rather than clinicians.

III. Methodology

A. Problem Formulation

The EHR data consist of a set of visit $V = \{v_1, v_2, \ldots, v_{|V|}\}$ for patient cohort $P = \{p_1, p_2, \ldots, p_i\}$; each visit v_i includes the patient's demographic information, Glasgow Coma Scale score, and a set of physiological features Fea, which are aggregated into a single record within 12-hour time windows $W = \{w_1, w_2, \ldots, w_t\}$. For each patient p_i, the medical ontology of all visits is composed of $\mathcal{C}$, $\mathcal{M}$, and $\mathcal{PR}$, where for each visit the condition set $C_i = \{c_1, c_2, \ldots, c_{|C_i|}\} \in \mathcal{C}$, the medication set $M_i = \{m_1, m_2, \ldots, m_{|m_i|}\} \in \mathcal{M}$, and the procedure set $\mathrm{PR}_i = \{\mathrm{pr}_1, \mathrm{pr}_2, \ldots, \mathrm{pr}_{|\mathrm{PR}_i|}\} \in \mathcal{PR}$. Given the clinical visit set V and the medical ontology $\mathcal{C}$, $\mathcal{M}$, and $\mathcal{PR}$, our goal is to predict in-hospital mortality and 30-day readmission (a binary classification task, where 1 denotes positive and 0 negative).

B. EHR Dataset Reconstruction

To refine EHR processing at finer granularity and mitigate data sparsity. First, we extract features from the dataset and anonymize patient data by shuffling patient order and reassigning indices. Next, we apply two imputation methods—Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB)—to handle missing values. Finally, the processed data are partitioned into features, ground-truth labels, and medical concepts for subsequent use.

TABLE I: Time-aware Score Calculation Table for Clinical Indicators

Features (Unit)	Score 0	Score 1	Score 2	Score 3
CRT < 3 s	✓			
CRT > 3 s			✓	
DBP (mmHg)	60–80	50–59 or 81–90	40–49 or 91–100	0–39 or 101–150
SBP (mmHg)	100–120	90–99 or 121–140	80–89 or 141–160	0–79 or 161–240
MBP (mmHg)	70–100	60–69 or 101–110	50–59 or 111–130	0–49 or 131–180
HR (bpm)	60–100	50–59 or 101–110	40–49 or 111–130	0–39 or 131–170
RR (bpm)	15–18	11–14 or 19–22	8–10 or 23–25	0–7 or 26–34
Temp (°C)	36.1–37.2	35.5–36 or 37.3–38	34–35.4 or 38.1–39.5	30–33.9 or 39.6–45
SpO_2 (%)	95–100	92–94	90–91	0–89
FiO_2 (%)	21–100	18–20	15–17	0–14
Gluc (mg/dL)	70–100	60–69 or 101–140	50–59 or 141–180	0–49 or 181–220
pH	7.35–7.45	7.3–7.34 or 7.46–7.5	7.2–7.29 or 7.51–7.6	7.0–7.19 or 7.61–7.8

C. Temporal Aware Score Computation & Sequential Model Training

To quantitatively assess the influence of feature values in different time windows on patients, we propose a scoring scheme inspired by related studies [17], as shown in Table I. A higher score indicates that the feature deviates further from the normal range and thus may have a greater impact. Next, we introduce the concrete computation of the time-aware score. For each patient visit, we extract physiological indicators per time window and assign each indicator a score from 0 to 3, reflecting its degree of deviation from normal.

Considering the dynamic nature of patient status, we adopt a multi-scale sliding window strategy, using various window lengths to aggregate scores within each time window. The time-aware score for a window is calculated as follows:

$$\mathrm{TAS\_Block}_t = \frac{1}{\kappa}\left(\omega \cdot \mathrm{GCS} + \frac{1}{m}\sum_{f=1}^{m} \mathrm{Fea\_mean}\right) \quad (1)$$

Here, GCS denotes the mean Glasgow Coma Scale within the current window (ranging from 3 to 15), and $\sum_{f=1}^{m} \mathrm{Fea\_mean}$ represents the sum of means of the other feature scores. The parameters κ and ω are scaling factors (set to $\kappa = 8$ and $\omega = 4/15$ in this work).

After obtaining the score sequence for all windows under record Fea, we use a pooling layer to map it to a fixed-length vector, thereby unifying dimensions across records. Accordingly, the time-aware score representation for patient p_i during visit v_i is:

$$\mathrm{TAS}_{v_i}^{p_i} = \mathrm{pooling}\Big(\{\mathrm{TAS\_Block}_k\}_{k=1}^{n};\, D\Big) \quad (2)$$

where D denotes the dimension of the pooled vector. Subsequently, we concatenate the time-aware score of each visit with the patient's feature record Fea as input to the model, which outputs the final prediction $\hat{y}_1$.

TABLE II: Dataset split statistics(Out. = in-hospital mortality, Re. = 30-day readmission).

Dataset	Split	$\text{Samples}_{\text{Out.}}$	$\text{Label}_{\text{Out.}}$	Avg $\text{Visit}_{\text{Out.}}$	$\text{Samples}_{\text{Re.}}$	$\text{Label}_{\text{Re.}}$	Avg $\text{Visit}_{\text{Re.}}$
MIMIC-III	Train	14000 (70%)	1844 (13.17%)	1.24	8750 (70%)	1882 (21.51%)	1.31
	Val	2000 (10%)	255 (12.75%)	1.24	1250 (10%)	264 (21.12%)	1.31
	Test	4000 (20%)	501 (12.53%)	1.25	2500 (20%)	511 (20.44%)	1.28
MIMIC-IV	Train	14000 (70%)	1815 (12.96%)	2.49	14000 (70%)	2347 (16.76%)	2.19
	Val	2000 (10%)	257 (12.85%)	2.45	2000 (10%)	345 (17.25%)	2.26
	Test	4000 (20%)	528 (13.20%)	2.55	4000 (20%)	708 (17.70%)	2.27

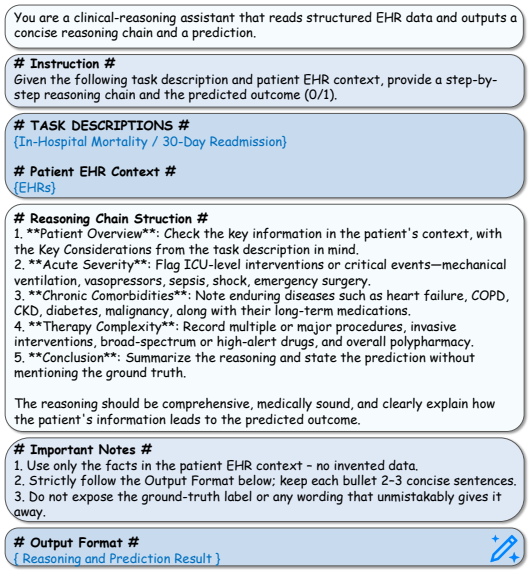

Fig. 3: Prompt template used by the local-reasoning LLM.

To model time-aware risk sequences across multiple visits, we design a deep sequential network comprising Bi-LSTM, self-attention, and an MLP. During training, we build input sequences on a per-patient basis. Because patients differ in the number of visits, sequence lengths vary; for each batch we first determine the maximum sequence length L_max and pad shorter sequences with zero vectors at the end, while constructing a mask matrix to mark valid and padded positions.

$$\text{MASK}_{p_i,j} = \begin{cases} 1, & \text{if position } j \text{ in patient } p_i \text{ is valid,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The mask is applied during LSTM encoding, attention pooling, and loss computation to ignore invalid positions, ensuring the model learns solely from genuine visits.

D. Local Reasoning LLM Training

To train a local LLM that can output reasoning chains and predictions based on patient medical concepts. First, we locally deploy a 32B-Qwen-2.5-instruct model and, given the patient's medical-concept context $\mathcal{C}$, $\mathcal{M}$, and $\mathcal{PR}$, use custom prompts to elicit the patient-specific chain of thought R_{p_i} . Erroneous or problematic outputs are removed through code-based filtering. We then randomly sample reasoning chains from 10,000 patients in the training set to fine-tune the local-Reasoning LLM. The instruction-tuning dataset is built with the patient's medical-concept context as input and the reasoning chain together with the ground-truth label as output.

$$D_{\text{LLM}} = \{(x_i,\, y_i)\}_{i=1}^{N} \tag{4}$$

Here, $x_i = \text{messages}_i = [\text{system prompt}, \text{user prompt}]$ denotes the input of the $i-th$ sample, comprising a system instruction and a user prompt derived from patient p_i's medical-concept context $\mathcal{C}$, $\mathcal{M}$, and $\mathcal{PR}$. $y_i = \text{response}_i$ denotes the LLM-generated reasoning chain together with the prediction result. Next, we fine-tune the local-Reasoning LLM (7B-Qwen-instruct) with D_{LLM}, augmenting the chain-of-thought prompt with instructions and formatted tags that guide the model to output binary predictions.

E. Co-prediction by sequence model and local-Reasoning LLM

To fuse the predictions of two models for optimal performance, we combine the output probabilities from the sequence prediction model and the local reasoning LLM predictor. Specifically, the sequence model first predicts each patient based on structured features Fea and the time-aware score, producing a probability $\hat{y}_1$. Then, the locally deployed language model generates a reasoning chain and prediction from the patient's medical-concept context $\mathcal{C}$, $\mathcal{M}$, and $\mathcal{PR}$, extracts the final prediction $\hat{y}_2$, and converts it to a float, with the prompt shown in Fig. 3. Finally, we adopt a weighted ensemble strategy to combine them into the final prediction probability:

$$\hat{y} = \text{step}(\alpha\hat{y}_1 + (1-\alpha)\hat{y}_2) \tag{5}$$

$$\text{step}(x) = \begin{cases} 1, & x \geq \beta, \\ 0, & x < \beta. \end{cases} \tag{6}$$

Here, $\alpha \in [0,1]$ denotes the ensemble weight, whose optimal value is found via grid search on the validation set. We then further search for the optimal classification threshold $\beta \in [0,1]$ on the validation set.

IV. Experiment

A. Experimental Setting

1) Datasets: We conduct experiments on the MIMIC-III and MIMIC-IV datasets, with their statistics presented in

TABLE III: In-hospital mortality and 30-day readmission prediction results on the MIMIC-III and MIMIC-IV datasets. **Bold** indicates the best performance. All metrics are multiplied by 100 for readability.

Methods	MIMIC-III Mortality			MIMIC-III Readmission			MIMIC-IV Mortality			MIMIC-IV Readmission		
	AUROC ↑	AUPRC ↑	min(+P,Se) ↑	AUROC ↑	AUPRC ↑	min(+P,Se) ↑	AUROC ↑	AUPRC ↑	min(+P,Se) ↑	AUROC ↑	AUPRC ↑	min(+P,Se) ↑
GRAM	84.70	49.21	49.64	77.84	47.97	46.95	87.75	54.01	54.62	79.53	50.13	50.80
KAME	84.59	49.48	49.51	78.04	48.23	47.41	87.76	55.74	54.79	78.91	47.62	49.63
CGL	84.20	47.64	47.67	77.47	46.68	47.73	88.42	56.64	54.80	78.95	47.74	49.16
MedPath	85.61	48.90	48.86	77.92	45.66	45.72	88.85	56.82	57.96	78.88	47.58	49.75
KerPrint	85.29	51.23	50.88	78.81	47.92	47.32	88.28	57.90	55.12	79.84	53.55	52.34
MPIM	85.24	50.52	50.59	78.65	48.26	46.94	89.45	60.10	57.62	79.13	47.67	49.52
MedGTX	85.97	49.36	48.20	75.60	46.44	45.99	88.77	58.33	58.25	78.82	47.48	49.54
GraphCare	85.85	50.16	49.15	78.70	47.19	46.82	89.13	60.85	59.16	79.18	48.55	49.64
MedRetriever	85.62	49.99	49.03	77.77	46.81	46.89	89.01	57.75	58.16	79.15	48.26	49.49
M3Care	83.33	47.86	49.96	76.80	46.29	45.38	88.14	54.06	54.30	79.87	51.03	51.10
UMM	84.01	49.76	49.41	77.46	47.81	47.24	87.82	53.84	55.40	78.75	48.63	49.58
VecoCare	83.43	47.28	47.92	76.93	46.18	47.22	88.01	55.37	55.35	79.17	51.58	54.42
EMERGE	86.25	52.08	51.42	**79.06**	48.59	47.86	89.50	63.11	59.95	80.61	57.28	54.50
EHR-TASR	**88.63**	**63.84**	**55.89**	76.70	**50.30**	**48.26**	**93.22**	**72.23**	**68.75**	**80.76**	**65.36**	**56.24**

Table II. MIMIC-III (v1.4)[2] is a large-scale, de-identified EHR dataset containing over 53,000 ICU admissions from 2001 to 2012, while MIMIC-IV (v2.2)[3] extends this resource to more than 380,000 admissions between 2008 and 2019.

2) Task: We focus on two EHR-based binary classification tasks: in-hospital mortality and 30-day readmission. The former uses all available visits to predict in-hospital death, while the latter identifies readmission within 30 days by scanning visits in reverse order and labeling patients accordingly.

3) Baselines: We compare our model with two categories of EHR modeling approaches: knowledge-enhanced graph methods and reasoning-augmented data-driven methods. The former integrates medical knowledge graphs to improve representation learning and interpretability, including GRAM [19], KAME [20], CGL [21], MedPath [22], KerPrint [23], MPIM [24], MedGTX [25], and GraphCare [18]. The latter leverages multimodal fusion, retrieval, or reasoning mechanisms to enhance generalization and clinical applicability, covering representative methods such as MedRetriever [26], M3Care [27], UMM [28], VecoCare [29], and EMERGE [8].

4) Implementation Details: In this study, Sliding-window lengths are set to 5, 10, and Full for MIMIC-III, and 3, 5, and Full for MIMIC-IV. The random seed is fixed at 42. The sequence model is trained with the binary cross-entropy (BCE) loss, a batch size of 128, and a learning rate of 0.001, using early stopping. The LLM is deployed via vLLM and fine-tuned with LlamaFactory using QLoRA for three epochs. Prompts are truncated to 7000 tokens and the model is quantized to int4. The optimal α and β are 0.79 and 0.57 for the outcome task, and 0.91 and 0.56 for the readmission task, respectively. All experiments are conducted on a Linux server with four NVIDIA RTX 3090 GPUs.

B. Main Experimental Results

Table III summarizes the performance comparison between EHR-TASR and 13 competitive baselines on MIMIC-III and MIMIC-IV. EHR-TASR consistently achieves the best or second-best performance across both in-hospital mortality and 30-day readmission tasks, outperforming state-of-the-art methods on most metrics, which demonstrates its strong effectiveness and robustness in clinical risk prediction. Notably, EHR-TASR yields substantial gains in AUPRC and min(+P, Se) on both datasets, highlighting its superior capability in identifying positive cases under severe class imbalance. This improvement suggests that dynamic time-aware representations are crucial for modeling patient disease trajectories and enhancing early risk discrimination. In addition, the locally deployed LLM not only contributes to improved predictive accuracy through fused decision making, but also provides clinically meaningful interpretability, while avoiding privacy risks associated with centralized medical data processing.

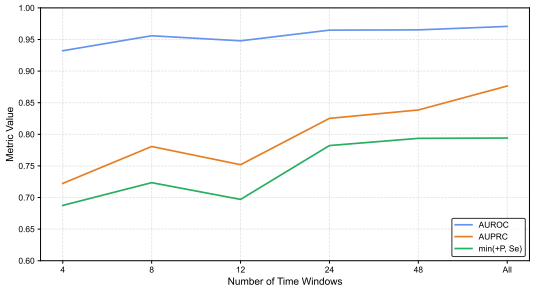

Fig. 4: Performance of Data with Different Sequence Lengths.

C. Ablation Study

Table IV demonstrates that each component of EHR-TASR contributes significantly to its overall performance. Notably, the sequential model is most critical for capturing longitudinal disease progression, as its removal causes the most severe performance drop. Ultimately, the full EHR-TASR framework achieves substantial improvements in AUPRC and min(+P, Se), particularly for the readmission task, confirming that fusing sequential modeling with LLM-based reasoning effectively enhances positive case identification under class imbalance.

D. Study on the Impact of Different Sequence Lengths

Figure 4 illustrates the performance of the model on the outcome task of the MIMIC-IV dataset when varying the number of time windows used for each patient. As the sequence

[2]https://physionet.org/content/mimiciii/1.4/
[3]https://physionet.org/content/mimiciv/3.1/

TABLE IV: Ablation study on the MIMIC-III and MIMIC-IV datasets.

Setting	MIMIC-III Mortality			MIMIC-III Readmission			MIMIC-IV Mortality			MIMIC-IV Readmission		
	AUROC	AUPRC	min(+P,Se)	AUROC	AUPRC	min(+P,Se)	AUROC	AUPRC	min(+P,Se)	AUROC	AUPRC	min(+P,Se)
EHR-TASR	**88.63**	**63.84**	**55.89**	**76.7**	**50.3**	**48.26**	**93.22**	**72.23**	**68.75**	**80.76**	**65.36**	**56.24**
w/o LLM	86.80	61.67	53.89	71.00	30.01	27.02	89.99	67.33	60.14	78.03	54.51	47.00
w/o LLM & TAS	86.01	61.33	50.52	64.88	26.48	26.26	88.94	64.33	55.81	76.01	52.50	46.08
w/o sequential model	59.33	32.22	24.29	54.58	22.96	17.81	73.06	42.46	24.29	54.69	20.21	16.95

length increases, all three metrics consistently improve. This trend can be attributed to the inclusion of richer longitudinal information, which provides a more comprehensive view of patient states and enables the time-aware module to more effectively capture fine-grained physiological variations.

V. Conclusion

In this paper, we propose EHR-TASR, which integrates a sequential model with a locally fine-tuned lightweight LLM for EHR-based risk prediction. EHR-TASR captures fine-grained temporal dynamics through multi-scale time-aware score computation, while leveraging LLM to generate structured reasoning aligned with the prediction outcomes, thereby enhancing interpretability and clinician usability. Extensive experiments on MIMIC-III and MIMIC-IV show that EHR-TASR consistently outperforms state-of-the-art methods. These results demonstrate that jointly modeling fine-grained temporal sequences and an interpretable LLM offers a powerful solution for reliable healthcare prediction.

References

[1] A. Tapuria, T. Porat, D. Kalra, G. Dsouza, X. Sun, and V. Curcin, "Impact of patient access to their electronic health record: A systematic review," *Inform. Health Soc. Care*, vol. 46, pp. 1–13, 2021.

[2] S. Wei, X. Peng, Y.-F. Wang, J. Si, W. Zhang, W. Lu, X. Wu, and Y. Wang, "BianCang: A Traditional Chinese Medicine Large Language Model," *arXiv preprint arXiv:2411.11027*, 2024.

[3] G. Huang, Y. Li, S. Jameel, Y. Long, and G. Papanastasiou, "From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?" *Comput. Struct. Biotechnol. J.*, vol. 24, pp. 362–373, 2024.

[4] Y. Li, Q. Zhang, W. Lu, X. Peng, W. Zhang, J. Si, Y. Gong, and L. Hu, "Time-aware Medication Recommendation via Intervention of Dynamic Treatment Regimes," in *WWW*, 2025.

[5] Z. Yu, C. Zhang, Y. Wang, W. Tang, J. Wang, and L. Ma, "Predict and interpret health risk using EHR through typical patients," in *ICASSP*, Seoul, South Korea, 2024, pp. 1506–1510.

[6] Y. Zhu, Z. Wang, L. He, S. Xie, X. Zheng, L. Ma, and C. Pan, "PRISM: Mitigating EHR data sparsity via learning from missing feature calibrated prototype patient representations," in *CIKM*, Boise, ID, USA, 2024, pp. 3560–3569.

[7] R. Xu, W. Shi, Y. Yu, Y. Zhuang, B. Jin, M. D. Wang, J. Ho, and C. Yang, "RAM-EHR: Retrieval augmentation meets clinical predictions on electronic health records," in *ACL*, 2024, pp. 754–765.

[8] Y. Zhu, C. Ren, Z. Wang, X. Zheng, S. Xie, J. Feng, X. Zhu, Z. Li, L. Ma, and C. Pan, "EMERGE: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation," in *CIKM*, Boise, ID, USA, 2024, pp. 3549–3559.

[9] R. AlSaad, A. Abd-alrazaq, S. Boughorbel, A. Ahmed, M.-A. Renault, R. Damseh, and J. Sheikh, "Multimodal large language models in health care: Applications, challenges, and future outlook," *J. Med. Internet Res.*, vol. 26, p. e59505, 2024.

[10] F. Yang, J. Zhang, W. Chen, Y. Lai, Y. Wang, and Q. Zou, "DeepMPM: A mortality risk prediction model using longitudinal EHR data," *BMC Bioinformatics*, vol. 23, no. 1, p. 423, 2022.

[11] M. Al Olaimat and S. Bozdag, "TA-RNN: An attention-based time-aware recurrent neural network architecture for electronic health records," *Bioinformatics*, vol. 40, Suppl. 1, pp. i169–i179, 2024.

[12] Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T. Lukasiewicz, and K. Rahimi, "Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 2, pp. 1106–1117, 2023.

[13] Z. Yang, A. Mitra, W. Liu, D. Berlowitz, and H. Yu, "TransformEHR: Transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records," *Nat. Commun.*, vol. 14, no. 1, p. 7857, 2023.

[14] G. O. Ghosheh, J. Li, and T. Zhu, "IGNITE: Individualized generation of imputations in time-series electronic health records," arXiv:2401.04402, 2024.

[15] W. Liao, Y. Zhu, Z. Zhang, Y. Wang, Z. Wang, X. Chu, Y. Wang, and L. Ma, "Learnable prompt as pseudo-imputation: Rethinking the necessity of traditional EHR data imputation in downstream clinical prediction," in *KDD*, Toronto, ON, Canada, 2025, pp. 765–776.

[16] T. Kwon, K. T. Ong, D. Kang, S. Moon, J. R. Lee, D. Hwang, Y. Sim, B. Sohn, D. Lee, and J. Yeo, "Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales," in *AAAI*, 2024, pp. 18417–18425.

[17] Y. Zhu, Z. Wang, J. Gao, Y. Tong, J. An, W. Liao, E. M. Harrison, L. Ma, and C. Pan, "Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data," arXiv:2402.01713, 2024.

[18] P. Jiang, C. Xiao, A. Cross, and J. Sun, "GraphCare: Enhancing healthcare predictions with personalized knowledge graphs," in *ICLR*, 2024.

[19] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *KDD*, 2017, pp. 787–795.

[20] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: Knowledge-based attention model for diagnosis prediction in healthcare," in *CIKM*, 2018, pp. 743–752.

[21] C. Lu, C. K. Reddy, P. Chakraborty, S. Kleinberg, and Y. Ning, "Collaborative graph learning with auxiliary text for temporal event prediction in healthcare," in *IJCAI*, 2021, pp. 3529–3535.

[22] M. Ye, S. Cui, Y. Wang, J. Luo, C. Xiao, and F. Ma, "MedPath: Augmenting health risk prediction via medical knowledge paths," in *WWW*, 2021, pp. 1397–1409.

[23] K. Yang, Y. Xu, P. Zou, H. Ding, J. Zhao, Y. Wang, and B. Xie, "KerPrint: Local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations," in *AAAI*, 2023.

[24] X. Zhang, S. Li, Z. Chen, X. Yan, and L. Petzold, "Improving medical predictions by irregular multimodal electronic health records modeling," in *ICML*, 2023.

[25] S. Park, S. Bae, J. Kim, T. Kim, and E. Choi, "Graph-text multi-modal pre-training for medical representation learning," in *CHIL*, 2022.

[26] M. Ye, S. Cui, Y. Wang, J. Luo, C. Xiao, and F. Ma, "MedRetriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text," in *CIKM*, 2021, pp. 2414–2423.

[27] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao, "M3Care: Learning with missing modalities in multimodal healthcare data," in *KDD*, 2022, pp. 2418–2428.

[28] K. Lee, S. Lee, S. Hahn, H. Hyun, E. Choi, B. Ahn, and J. Lee, "Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention," in *MLHC*, PMLR, vol. 219, pp. 423–442, 2023.

[29] Y. Xu, K. Yang, C. Zhang, P. Zou, Z. Wang, H. Ding, J. Zhao, Y. Wang, and B. Xie, "VecoCare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data," in *IJCAI*, 2023, pp. 4921–4929.