Decode Like a Clinician: Enhancing LLM Fine-Tuning with Temporal Structured Data Representation

Daniel Fadlon[1,#], **David Dov**[2,#], **Aviya Bennett**[2], **Daphna Heller-Miron**[2], **Gad Levy**[2], **Kfir Bar**[1,*], **Ahuva Weiss-Meilik**[2,*]

[1]Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
[2]Project M

Correspondence: kfir.bar@runi.ac.il

Abstract

Predictive modeling of hospital patient data is challenging due to its structured format, irregular timing of measurements, and variation in data representation across institutions. While traditional models often struggle with such inconsistencies, Large Language Models (LLMs) offer a flexible alternative. In this work, we propose a method for verbalizing structured Electronic Health Records (EHRs) into a format suitable for LLMs and systematically examine how to include time-stamped clinical observations—such as lab tests and vital signs—from previous time points in the prompt. We study how different ways of structuring this temporal information affect predictive performance, and whether fine-tuning alone enables LLMs to effectively reason over such data. Evaluated on two real-world hospital datasets and MIMIC-IV, our approach achieves strong in-hospital and cross-hospital performance, laying the groundwork for more generalizable clinical modeling. Code and models are available at `https://github.com/DanielFadlon/decode-like-a-clinician`.

1 Introduction

Electronic medical records (EMRs) used in hospitals are inherently complex, comprising both structured data—such as demographic details and lab results—and unstructured content like clinical notes and diagnostic reports. Moreover, EMRs exhibit sparsity and temporal irregularity, as medical events and measurements are recorded intermittently, reflecting the progression of the patient's condition and the timing of assessments, treatments, and other clinical procedures.

Consider a patient admitted to an internal medicine ward. Throughout the hospitalization, the medical team periodically orders tests—such as blood work or imaging—and continuously monitors certain parameters like blood pressure or heart rate. These data points are not recorded at uniform intervals; rather, they arrive asynchronously, depending on clinical decisions, test processing times, and the patient's evolving condition. Some measurements, like vital signs, may fluctuate rapidly, while others, such as lab values, change more gradually. This presents a major challenge for training machine learning models—such as XGBoost (Chen and Guestrin, 2016) and neural networks—for clinical outcome prediction, as these models typically rely on regular, complete, and temporally aligned inputs. Although XGBoost is widely used in medical prediction tasks, it struggles with heterogeneous data, cross-hospital generalization, and the sparsity and temporal irregularities characteristic of real-world patient trajectories.

These challenges often require extensive data preprocessing and constrain the models' ability to fully leverage the rich and dynamic information available throughout hospitalization.

Recent advancements in generative large language models (LLMs) present a promising avenue for addressing these challenges. LLMs, with their inherent flexibility and proficiency in processing textual information, offer the potential to learn from rich textual representations. This approach not only simplifies the data preprocessing pipeline, but also enhances the model's ability to integrate diverse data types, and develop a more comprehensive understanding of the patient's journey.

In this work, we present a pipeline that encodes the temporal aspects of structured EMRs into the prompt for an LLM, followed by fine-tuning step to predict a real-world clinical outcome.

We study the impact of different temporal aggregation strategies, the role of time annotations (absolute and relative timestamps), and the effect of incorporating larger volumes of patient history on prediction quality. We also explore whether

[*] These authors jointly supervised this work.
[#] The authors contributed equally to this work.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1906–1922
December 20-24, 2025 ©2025 Association for Computational Linguistics

LLMs can generalize across hospitals and patient trajectories. Using real-world data from two leading hospitals and the publicly available MIMIC-IV dataset (Johnson et al., 2022), we evaluate the generalizability of our method across varied clinical settings and demonstrate their effectiveness in predicting clinical outcomes.

Our results show that aggregating time-sensitive clinical events into the LLM prompt significantly improves performance during fine-tuning for clinical outcome prediction, highlighting the model's ability to capture temporal patterns in clinical records.

To summarize, this work makes the following contributions: (1) We propose a novel method for encoding temporal EMRs into prompts for LLMs, enabling effective fine-tuning for downstream clinical prediction tasks. Our approach achieves strong performance in both in-hospital and cross-hospital evaluations, validated on real-world and open-source datasets. (2) We conduct a comprehensive analysis of the strengths and limitations of an LLM in modeling the temporal structure of EMRs, examining how different event verbalization strategies influence predictive performance.

2 Related Work

Early findings suggest that pre-trained LLMs remain less reliable than specialized models for structured clinical prediction tasks. The CLINICALBENCH (Canyu Chen, 2024) shows that serializing EHRs into text and querying foundation models such as GPT-4 (OpenAI, 2024) and Llama-3 (Dubey et al., 2024) significantly underperforms compared to gradient-boosted trees and LSTMs on tasks like mortality prediction and length-of-stay estimation. In a head-to-head comparison on delirium prediction (Mohamed Rezk and Dahlweid, 2024), GPT-4 missed 38% of true-positive cases relative to the proprietary *Clinicalytix* neural network. Similarly, Gao et al. (2024) report that XGBoost trained on raw tabular data consistently outperforms a range of LLM-based embedding pipelines.

At the same time, a growing body of research highlights the promise of LLMs when structured data is carefully encoded. Hegselmann et al. (2023) (TABLLM) show that with carefully designed prompts, few-shot GPT-3 (Brown et al., 2020) can outperform deep tabular baselines on small-scale classification tasks. For relational data, the TALK LIKE A GRAPH framework (Bahare Fatemi, 2024) demonstrates that effective graph verbalization significantly enhances LLM reasoning. LLMs are also showing emerging competence in modeling temporal structure: when prediction is framed as a next-token generation task, GPT-3 and Llama-2 (Touvron et al., 2023) perform comparably to classical time-series models (Gruver et al., 2023), while TABPFN further improves performance by incorporating explicit temporal annotations (Hollmann et al., 2025).

Fine-tuning emerges as a promising alternative when zero-shot LLMs fall short on clinical outcome prediction tasks—such as in-hospital mortality or sepsis detection. Yet, the role of structured-data representation (i.e., table-to-text) in the fine-tuning process remains underexplored, especially in non-ICU hospital settings where data sparsity presents a major challenge.

While LLMs excel at processing text, their performance on structured data is often less robust. Most recent efforts focus on combining structured EHR data with textual inputs (e.g., clinical notes) to improve predictive performance. For example, HEALTH-LLM (Yubin Kim, 2024) leverages context-rich prompts incorporating user profiles and temporal cues to adapt LLMs to wearable sensor data. Unlike their emphasis on continuous data streams, we address the challenges of sparse, irregular clinical events. Alba (2025) finetune BioClinicalBERT(Alsentzer et al., 2019) and BioGPT (Renqian Luo, 2022) on clinical notes, demonstrating that textual summaries and temporal cues enhance predictions of 30-day mortality and Deep Vein Thrombosis (DVT). Battula et al. (2024) integrate LLM-generated ”expert summaries” of ICU notes with structured time-series data, using the 70B-parameter Med42-v2 model (Christophe, 2024). Similarly, Supreeth P. Shashikumar (2025) show that LLM-generated text can reduce false alarms by over 50% in emergency department sepsis models. In a related work, Naik et al. (2022) propose BEEP, a system that retrieves patient-specific medical literature to augment clinical notes for predictive modeling.

A smaller body of work focuses exclusively on structured data. LLAMACARE (Li et al., 2024) uses GPT-4-generated summaries to enrich clinical features and instruction-tunes Llama-2 via LoRA. Likewise, CPLLM (Shoham. and Rappoport., 2025) leverages LLMs to verbalize structured inputs before fine-tuning with binary labels. Although these methods outperform baseline LLMs such as PMC-LLaMA (Wu C, 2024) and LLama-2 Chat, they lack benchmarking against robust tabular models such as XGBoost. Moreover, LLAMACARE's human evaluation suggests that LLM-generated data is useful but still limited. To the best of our knowledge, no study has demonstrated that LLMs trained solely on clinical tabular data outperform XGBoost in outcome prediction tasks. Together, these works highlight the importance of data representation in LLM fine-tuning—whether through verbalization or data integration. However, none of them systematically evaluate how different representation strategies affect model performance or whether fine-tuning alone can rival traditional tabular models. Our work directly addresses this gap. We systematically evaluate LLMs' ability to learn and reason over verbalized structured data using various prompting strategies in sparse, non-ICU hospital settings. We also propose a straightforward yet effective framework for verbalizing temporal structured hospital data.

3 Methodology

We address the task of clinical outcome prediction (e.g., mortality) using structured temporal data derived from patient admissions.

3.1 Structured Temporal Dataset

The dataset consists of structured clinical records, organized by individual patient admissions. Each admission is represented by a table that captures the patient's clinical state over time. The table is structured with rows representing consecutive, non-overlapping six-hour intervals—an approach commonly used in non–intensive care unit (non-ICU) datasets—and columns corresponding to clinical parameters (e.g., diastolic blood pressure, hemoglobin levels). Therefore, the first row of each admission represents the values of parameters measured during the first 6 hours of admission, the second row from 6 to 12 hours, and so on, each summarizing a distinct six-hour interval.

Each parameter may be measured multiple times within a six-hour interval. To consolidate these measurements into a single value per parameter per interval, we apply an aggregation function—either the maximum or the average of the values—depending on the specific dataset configuration. For example, if diastolic blood pressure (DBP) is measured three times during the first six-hour window with values 70, 75, and 80, and the aggregation function is *max*, the resulting value for that interval will be 80.

Each patient admission is also associated with a single binary label indicating whether a predefined clinical outcome (e.g., mortality, length of stay in the hospital) will occur at any point in the future during the current admission. This label serves as the prediction target.

To construct instances for classification, we segment each admission into snapshots. Each snapshot includes the data from the beginning of the admission (i.e., interval 0) up to a specified cutoff time, where the cutoff time t increases in multiples of 6 hours (e.g., 6h, 12h, 18h). Each snapshot is treated as an independent instance for training or evaluation purposes. So, for each patient admission that contains T time slots, we generate T snapshot instances, each using a different interval as its cutoff (see Figure 8 in Appendix A).

3.2 Verbalizer

Since LLMs require textual input, we introduce a verbalizer that converts each structured snapshot into natural language, enabling LLM-based training and inference.

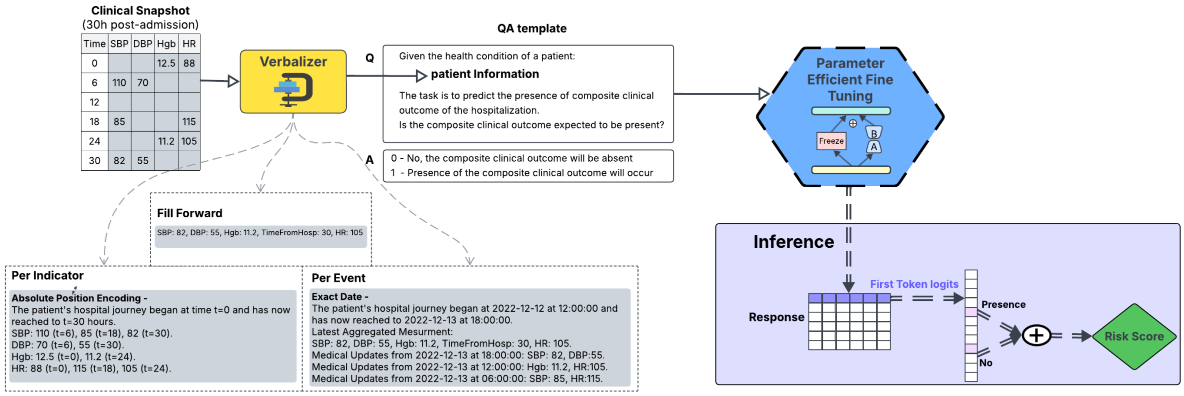
The goal of the verbalizer is to faithfully express the content of each snapshot—including the measured clinical parameters, their values, and the time intervals in which they were recorded—in a way that can be interpreted by the LLM. To this end, we explore several prompting strategies, each differing in how they organize temporal information, refer to parameter names, and present the values. By designing and comparing multiple verbalization schemes, we aim to evaluate how different prompt formats influence model performance and to identify which approaches best support prediction based on structured temporal clinical input.

To systematically explore the space of possible prompts, we define four key dimensions of variation in our verbalization process:

3.2.1 Data Aggregation Strategy

This dimension determines how values are grouped and presented in the prompt. We have three approaches:

Forward-Fill (FF). A common method for handling sequential clinical data, where each parameter is represented by its most recent observed value at each time snapshot. For example, if hemoglobin was measured sometimes during hour 6 and not



Clinical Snapshot
(30h post-admission)

Time	SBP	DBP	Hgb	HR
0			12.5	88
6	110	70		
12				
18	85			115
24			11.2	105
30	82	55		

QA template

Q: Given the health condition of a patient:
patient Information
The task is to predict the presence of composite clinical outcome of the hospitalization.
Is the composite clinical outcome expected to be present?

A: 0 - No, the composite clinical outcome will be absent
1 - Presence of the composite clinical outcome will occur

Fill Forward
SBP: 82, DBP: 55, Hgb: 11.2, TimeFromHosp: 30, HR: 105

Per Indicator
Absolute Position Encoding -
The patient's hospital journey began at time t=0 and has now reached to t=30 hours.
SBP: 110 (t=6), 85 (t=18), 82 (t=30).
DBP: 70 (t=6), 55 (t=30).
Hgb: 12.5 (t=0), 11.2 (t=24).
HR: 88 (t=0), 115 (t=18), 105 (t=24).

Per Event
Exact Date -
The patient's hospital journey began at 2022-12-12 at 12:00:00 and has now reached to 2022-12-13 at 18:00:00.
Latest Aggregated Mesurment:
SBP: 82, DBP: 55, Hgb: 11.2, TimeFromHosp: 30, HR: 105.
Medical Updates from 2022-12-13 at 18:00:00: SBP: 82, DBP:55.
Medical Updates from 2022-12-13 at 12:00:00: Hgb: 11.2, HR:105.
Medical Updates from 2022-12-13 at 06:00:00: SBP: 85, HR:115.

Figure 1: Pipeline – Given a clinical snapshot of a patient's condition at time t hours from admission, the Verbalizer component aggregates and formats the data. This processed information is then wrapped into a structured QA template and used to fine-tune a large language model (LLM). During inference, the same verbalization method is applied to raw input data to generate a prompt for the fine-tuned LLM. The model's output is used to compute a risk probability, obtained by extracting the logits of the first generated token and normalizing over the vocabulary probabilities of the predefined outcome tokens: "No" (0) and "Presence" (1).

again until hour 24, then for all intermediate time slots (e.g., 12 and 18), the value from interval 6 would be used as the most recent observation.

Per-indicator. The prompt groups all measurements of a single clinical parameter, highlighting how it evolves over time. For example, *Hemoglobin: 13.2 at hour 6, 12.7 at hour 18, and 13.0 at hour 30.* This format emphasizes the trajectory of a single indicator across multiple time points. We adopt this representation to reflect the natural time-series structure of clinical data and to focus the model's attention on the progression of individual indicators—an approach that may be particularly effective when the temporal pattern of a specific parameter is clinically meaningful. The exact format of the prompt may vary depending on other dimensions of the verbalizer, such as the choice of time annotation and narrative style, which we describe below.

Per-event. The prompt focuses on co-occurring values—that is, all clinical parameters recorded together within the same time interval—capturing their joint configuration. For example, a prompt based on a single six-hour interval might read: *At hour 12, the patient's hemoglobin was 12.7, diastolic blood pressure (DBP) was 70, and glucose was 140.* We adopt this representation because it mirrors the way clinicians naturally assess patient status—by interpreting multiple measurements in context—and allows the model to attend to meaningful joint updates across parameters. Additionally, since clinical events are often documented in this format, it simplifies preprocessing and aligns closely with the native structure of the data.

3.2.2 Time annotation

Each value or group of values may be annotated with a time reference. The goal is to evaluate whether different representations of time affect the model's ability to integrate clinical information and accurately predict the clinical outcome. We experiment with several temporal formats:

No time. Values of parameters are listed without any reference to time.

Absolute time from admission. Each observation is marked by its offset from admission time. For example, Hemoglobin: 12.5 ($t = 12$), 11.2 ($t = 24$).

Relative time from cutoff. Time is expressed relative to the current moment (e.g., *Hemoglobin: 12.5 (24 hours ago), 11.2 (36 hours ago)*).

Exact timestamps. Inspired by prior work (Shi Bin Hoo, 2025), measured values are annotated with full timestamps (e.g., "2022-12-13 at 18:00:00"). Because the original timestamps were removed during the de-identification process, we standardize each admission to begin at 2022-12-02.

3.2.3 History Length

This dimension defines how much past information is included in the prompt. In the **per-indicator** mode, it controls how many previous values of each parameter are included. For example, if the history length is set to 2, the prompt for Hemoglobin might read: *Hemoglobin: 13.2 (t=6), 12.7 (t=18).* In the

per-event mode, it specifies how many past events are verbalized into the prompt. For example, with a history length of 2, the prompt might include:
Medical Updates from t=12: Heart Rate: 78, Diastolic Blood Pressure: 70.
Medical Updates from t=18: Hemoglobin 13.0.

Notably, with the *per-indicator* approach and a history length of 2, we can incorporate more complete historical information for each indicator. In contrast, the *per-event* setting limits the historical context for a given parameter to the two most recent recorded events.

3.2.4 Narrative Style

We vary the tone and structure of the text used to describe the snapshot:

Technical. Compact and schema-driven descriptions (e.g., *Hemoglobin at t=12: 13.2*).

Descriptive Richer, more narrative forms (e.g., *Twelve hours into the admission, the patient's hemoglobin level was measured at 13.2 grams per deciliter*). The procedure for constructing descriptive prompts is outlined in detail in Appendix D.

Each snapshot can thus be rendered in multiple ways depending on the configuration along these axes. We provide several examples in Appendix G.

3.3 Fine-Tuning and Inference

We use a generative LLM to predict clinical outcomes from verbalized snapshots of patient admissions. Our experiments focus on the fine-tuning setting, where the LLM is fine-tuned using text prompts generated by the verbalizer as input, paired with a corresponding output text representing the prediction. We also refer to this step as *instruction tuning*.

To guide the model toward generating structured responses, we append a final paragraph to each prompt, as illustrated in Figure 1, which frames the task as a question. This helps elicit a more natural and consistent response format from the model. Since all prediction tasks in our setup are binary classification, we map the two possible outcomes to the following canonical textual responses:

- **0**: *"No, the composite clinical outcome will be absent."*
- **1**: *"Presence of the composite clinical outcome will occur."*

During fine-tuning, the model is optimized to predict the next token in a sequence. At inference time, to determine the model's prediction, we analyze the token-level probability distribution of the first generated token and compare the probability of the tokens *presence* and *no*. The final label is assigned based on the token with the higher probability. We normalize the probabilities of the two target tokens to obtain a standard binary classification distribution, allowing the computation of probability-based evaluation metrics as described in Section 4.

3.4 Datasets

Our study is based on three datasets of patient admission electronic health records (EHRs). Two datasets were collected from two different medical centers in Israel as part of an ongoing collaboration: Tel Aviv Sourasky (Ichilov) Medical Center (**TASMC**) and Sheba Medical Center (**ShibaMC**). The third dataset is **MIMIC-IV**, a publicly available resource. All data usage was approved by the relevant institutional review boards, and all records were de-identified prior to analysis. For all three datasets, we extracted a structured set of clinical parameters, including demographics, laboratory results, vital signs, medication orders, and procedures.

While there is some overlap, the specific set of parameters varies across the three datasets, reflecting differences in clinical practices and data recording protocols.

The **TASMC** cohort comprises 7,004 admissions from 6,679 unique patients, recorded between 1 February 2014 and 30 June 2021. We partition the data into 5,623 admissions for training, 693 for validation (all on or before 31 May 2020), and 688 for testing (on or after 1 June 2020). To avoid data leakage, each patient appears in only one partition. For computational efficiency, we randomly sample 20% of the training set for model training while preserving the label ratio.

The **ShebaMC** dataset was extracted using the same pipeline, contributing an additional 1,388 admissions collected between 2021 and 2023. This cohort is used exclusively for cross-hospital evaluation to assess model performance on out-of-distribution (OOD) data.

Both hospital datasets include patients admitted with hip fractures and share the same prediction task, allowing controlled cross-hospital evaluation. Admissions are divided into six-hour intervals, with each clinical parameter aggregated using the maximum value within each interval. The clinical outcome is defined as a composite event including in-hospital mortality, transfer to an intensive care unit (ICU) or step-down unit (SDU), administration of inotropic medications, readmission within 30 days, re-operation, or a hospital length of stay (LOS) exceeding 15 days. An interval is labeled as 1 if any of these events occur later during the same admission.

MIMIC-IV (Johnson et al., 2022; Goldberger et al., 2000). We use the standardized, publicly available MIMIC-IV-Data-Pipeline repository[#] to preprocess the dataset (Gupta et al., 2022). Our pipeline largely follows the original extraction procedures, including feature selection, time interval definitions, and labeling strategies, with only minimal modifications. Specifically, we define a unified clinical outcome by combining three predefined tasks from the original pipeline—in-hospital mortality, heart failure in 30 days, and extended length of stay (LOS > 15 days). This composite outcome was designed to align MIMIC-IV with the datasets of the hospitals, facilitating a more comparable experimental setup. Mortality events are considered from the first observable event onward, prolonged LOS assessments are tracked starting day two, and heart failure diagnoses are evaluated starting day three to accommodate variability in phenotyping and the coexistence of acute and chronic conditions. Consequently, our deterioration task formulation captures a broader, clinically diverse representation of patient deterioration.

We use only the first 15 days of hospitalization for each admission, as admissions exceeding this duration are, by definition, assigned a positive outcome label. The data is aggregated into four-hour intervals using the mean of the observed values, following the methodology established in the MIMIC-IV-Data-Pipeline.

Like with the other two datasets, we partitioned the data chronologically into distinct training, validation, and test sets, ensuring that each patient appears in only one set to prevent data leakage. This process yields 808,183 instances derived from 33,618 admissions. However, to ensure computational feasibility, we use only 5% of these instances for training.

Table 1 summarizes the sizes of the three datasets we work with in this study, and how we split them into train/validation/test.

Dataset	Split	# Admissions	# Instances
TASMC	Train (all)	5,623	188,155
	Train	5,583	45,159
	Validation	693	20,242
	Test	688	19,148
ShebaMC	Train (all)	11,444	248,528
	Train	10,274	49,706
	Validation	1,145	30,953
	Test	1,388	36,072
MIMIC-IV	Train (all)	29,751	808,183
	Train	19,988	40,409
	Validation	1,788	50,542
	Test	2,079	56,000

Table 1: Dataset sizes. *Train(all)* refers to training on the full dataset, while *Train* refers to the actual subset training data we used for the generative models.

4 Experimental Settings

For all experiments, we fine-tuned the `LLaMa-3-8B-Instruct` model on the training set and evaluated on both validation and test sets. Training ran for up to five epochs with early stopping based on validation performance measured at the end of each epoch. Each experiment was repeated with three random seeds for robustness. See Appendix B for training details.

4.1 Evaluation

Our primary evaluation considers all snapshots per admission, reflecting real-world scenarios where predictions are made throughout hospitalization. Furthermore, we assess performance at specific time points by selecting a single snapshot—corresponding to the relevant interval—per admission (see Appendix F).

We evaluate model performance using ROC-AUC, as it effectively captures the model's discriminative ability across all thresholds. Probabilistic outputs are derived from the first generated token (see Figure 1). Initial evaluations are conducted on the test split (in-hospital). Top-performing configurations further assessed on the independent ShebaMC dataset to evaluate out-of-hospital generalization.

4.2 Baseline

Using the Forward-Fill method for our baselines, we include XGBoost as a strong benchmark due to its success with tabular clinical data, consistent with Gao et al. (2024) and Canyu Chen (2024). The XGBoost model is trained on the entire training set. Key hyperparameters are tuned on the validation set.

[#] https://github.com/healthylaife/MIMIC-IV-Data-Pipeline

Model	**Aggregation Method**	**Time Annotation**	**History Length**	**AUC-ROC** (mean ± std)	
				TASMC	**MIMIC-IV**
Baselines					
Me-LLaMA (ZSL)	Forward-Fill			58.26	62.34
XGBoost	Forward-Fill			76.03	84.09
Logistic-Regression	Forward-Fill			75.31	81.83
LSTM	(see the text)			73.78	84.19
LLaMA-3-8B -Instruct **(ours)**	Forward-Fill			75.82±1.14	84.13±0.37
	Per Indicator	No time	3	75.30±0.16	84.18±0.51
		Relative time	3	76.66±0.48	84.72±0.04
		Absolute time	2	76.80±0.12	84.23±0.16
			3	77.80±0.17	84.93±0.12
			4	77.93±0.33	85.28±0.12
		Exact timestamps	3	76.69±0.47	84.54±0.27
	Per Event	No time	3	76.02±1.05	87.48±0.06
		Relative time	3	76.33±0.40	88.06±0.22
		Absolute time	3	76.87±0.37	88.16±0.27
		Exact timestamps	3	77.06±0.23	88.23±0.15
			6	77.94±0.31	89.76±0.19
			8	**77.97**±0.33	**90.26**±0.06

Table 2: Main results.

To ensure comprehensive coverage of modeling paradigms, we further include a Logistic Regression baseline representing a simple linear model, and a Long short-term memory recurrent neural network (LSTM) baseline capturing temporal dependencies in patient trajectories. The LSTM model aggregates information from the last five recorded time points for each indicator. As the input is a structured table, all indicators are temporally aligned. For additional training details and the final used hyperparameter configuration see Appendix B

Finally, to assess the necessity of fine-tuning—we use `Me-LLaMA-13B` (Xie, 2024) (Xie et al., 2024), a medical LLM fine-tuned for clinical and biomedical text. Pre-trained on domain-specific data, including MIMIC-IV, it outperforms other open-source models in both zero-shot and supervised settings. We evaluate it in a zero-shot setting; full prompt and evaluation details are provided in Appendix C.

5 Results

The results are reported in Table 2. We organize the table such that each column reflects a specific component of the prompting strategy described in Section 3.2. Following the subpar performance of Me-LLaMA zero-shot learning compared to XGBoost and our fine-tuned models, we conclude that fine-tuning is necessary for these tasks, and we focus our analysis comparing to the XGBoost baseline. Below, we highlight several key observations drawn from the results.

Aggregation Strategy. Encoding lab results and medications using *per-indicator* or *per-event* formats outperforms the *forward-fill* strategy used by XGBoost and Llama-3-8B-Instruct. The *per-event* approach yields the best results (TASMC: 77.97 ± 0.03; MIMIC-IV: 90.26 ± 0.04), demonstrating the value of capturing fine-grained temporal structure. In contrast, fine-tuning an LLM with the *forward-fill* method does not surpass XGBoost trained on the same input, instead yielding similar performance. This highlights the critical role of data representation in the prompt for the fine-tuning process.

On the TASMC dataset, *per-indicator* and *per-event* performed similarly, while on MIMIC-IV, *per-event* outperformed *per-indicator* by 5.33% (90.26 vs. 84.93). This performance gap, as concluded from the label-specific analysis in Appendix E, emphasizes the effectiveness of the *per-event* aggregation method for heart failure prediction—a label exclusive to the MIMIC-IV task.

Time Annotation. Annotating events with the *absolute* hour improves AUC-ROC by 2.5 percentage points over the *no-time* approach in the *per-indicator* setting (75.30 vs. 77.80) and consistently outperforms *relative-time*. While time annotation plays a key role in *per-indicator* aggregation, its impact is less pronounced in the *per-event* setting.

History Length. Performance improves rapidly up to an inflection point—three indicator windows or six event windows—after which gains level off.

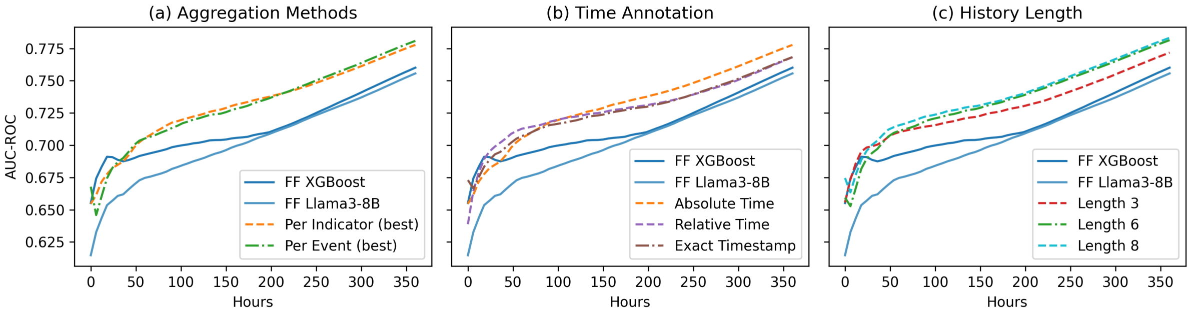


Figure 2: **TASMC AUC-ROC Analysis.** The figure compares AUC-ROC scores over time, with the x-axis showing hours and the y-axis showing cumulative AUC-ROC from time 0 to t. (a) Aggregation method performance on TASMC. (b) Time annotation results for *per-indicator* on TASMC. (c) History length analysis for *per-event* on TASMC.

Notably, longer input histories do not degrade performance, indicating the model's ability to scale with added clinical context up to a certain threshold.

In summary, combining *per-event* aggregation, *exact-timestamps* time annotation, and a moderate history window yields consistent gains and offers practical guidance for EHR representation in LLM fine-tuning.

5.1 Narrative Style

The descriptive approach—recasting technical input into a more descriptive clinical style—had mixed results. It slightly lowered performance on the TASMC dataset (-0.3 pp vs. the technical approach) but improved results in the MIMIC-IV setting (+0.7 pp) (full results are provided in Appendix D). We therefore remain inconclusive about the overall benefit of using descriptive prompts.

5.2 Cross Hospital Results

Table 3 summarizes the cross-hospital results, showing that LLM-based models generalize better than XGBoost. They effectively handle inconsistencies in feature naming and units, with the *per-event* representation achieving 74.67% AUC-ROC on ShebaMC, outperforming the *per-indicator* approach (73.03%) and XGBoost (68%).

While XGBoost trained in-hospital reaches 80.02%, the best fine-tuned LLM achieves 81.03%, showing that cross-hospital performance remains strong despite a drop from the top in-hospital result. Due to XGBoost's need for a fixed feature set, cross-hospital evaluations used only intersected features (30). LLMs, by contrast, maintained strong performance even with the full feature set (40), demonstrating greater flexibility and robustness to feature mismatches.

Experiment	**ShebaMC**		
	IF	**AF**	
	cross-hosp	*in*-hosp	*cross*-hosp
Baselines			
Me-LLaMA (ZSL)	59.43		57.80
XGBoost (FF)	68.02	80.02	
LLaMA (ours)			
Forward-Fill	70.88	79.94	72.88
Per Indicator *best*	73.03	**81.03**	72.97
Per Event *best*	**74.67**	80.93	**74.41**

Table 3: Cross-hospital results using **intersected features** (IF) and **all features** (AF). ZSL = Zero-shot learning. ***cross-hosp***: trained on TASMC, evaluated on ShebaMC. ***in-hosp***: trained and evaluated on ShebaMC.

5.3 Time-Based Performance Evaluation

To assess robustness over time, we evaluate cumulative performance at different time intervals, integrating data up to each point t. As shown in Figure 2, the model consistently improves across all intervals. While our main results use a 15-day cutoff, performance remains strong from 1 day (24 hours) to 15 days (360 hours).

6 Discussion

Our findings show that representation design determines how effectively LLMs reason over structured temporal data. The success of the *per-event* strategy highlights that aligning co-occurring observations within temporally grounded textual units enables LLMs to capture the dynamics of clinical trajectories without losing the precision of tabular input.

It is also instructive to consider how other domains handle similar challenges when translating structured or temporal information into language that LLMs can process. We group existing approaches into three main categories and position our work in relation to them.

Structured-based approaches [e.g., JSON or key-value linearization (Gao et al., 2024; Gupta et al., 2023; Shankarampeta et al., 2025)] rely on the assumption that data are static and complete—where each record represents a full, consistent snapshot of the world. These methods preserve schema fidelity and interpretability but ignore the temporal relationships between observations, making them less suited for dynamic, partially observed processes such as patient trajectories.

Temporal-aware approaches [e.g., TIME-LLM, LLMS ARE ZERO-SHOT TIME SERIES FORECASTERS, AAD-LLM (Jin et al., 2024; Gruver et al., 2023; Russell-Gilbert et al., 2024)] assume dense and regularly sampled time series. They aggregate information within fixed windows or deterministic statistics (mean, max, z-score), producing compact summaries suitable for continuous signals such as sensors or financial feeds. However, these assumptions break down in settings like EHRs, where updates are sparse, asynchronous, and event-triggered rather than uniformly sampled.

Narrative-based approaches [e.g., PROMPTCAST, HEALTHLLM (Xue and Salim, 2023; Yubin Kim, 2024)] assume that the underlying data stream is regular and coherent enough to be verbalized fluently as a story (e.g., "The temperatures are 77, 68, and 66. . . "). While this style improves interpretability, it sacrifices quantitative precision and temporal anchoring—two aspects essential in irregular, multi-sourced hospital data.

Our work lies at the intersection of these paradigms. We preserve structure through explicit event anchoring while maintaining linguistic coherence that facilitates reasoning. This *event-centric verbalization* framework is particularly suited to domains where events occur asynchronously and at low frequency—such as long-term hospitalizations or periodic customer support interactions—where each update carries substantial semantic information. We believe the approach can generalize to such settings, though its behavior in high-velocity environments remains an open question. In domains characterized by rapid, continuous updates, such as intensive care monitoring or financial markets, it will be important to examine whether simply updating temporal resolution or context window length suffices, or whether new aggregation and alignment mechanisms are needed to capture fine-grained temporal dynamics.

7 Conclusion

This work presents the first comprehensive evaluation of LLMs for interpreting sparse, temporally structured clinical data. By systematically comparing verbalization strategies, we identify a prompt configuration—*per-event* aggregation combined with *exact-timestamps* time annotation—that consistently enhance fine-tuning performance across different datasets and clinical tasks. These findings emphasize the importance of structured data representation in enabling LLMs to predict clinical outcome over longitudinal records.

To the best of our knowledge, this is also the first study to evaluate and compare cross-hospital generalization using structured clinical data. We show that with well-designed verbalization strategies, LLMs can match or even exceed traditional models like XGBoost in out-of-hospital settings.

Overall, our work provides a foundation for integrating structured data into LLM-based clinical pipelines. We encourage future research to build on these findings by expanding verbalization methods, incorporating multimodal inputs, and evaluating across broader clinical prediction tasks.

Limitations

This study has several limitations. First, we utilize a single pipeline for retrieving and preprocessing data from the MIMIC-IV dataset (Gupta et al., 2022). While this ensures consistency across experiments, it may limit the generalizability of our findings, as different preprocessing strategies or cohort definitions could yield different results.

Second, we conduct all experiments using a single model. This decision stems from the resource-intensive nature of fine-tuning large language models across multiple configurations. Given the substantial computational and financial cost involved, extending the study to include additional models was beyond our current scope, though it remains an important direction for future work.

Third, our method relies on supervised fine-tuning, which requires labeled data. While obtaining high-quality annotated datasets in clinical settings can be challenging, many real-world clinical outcomes are routinely recorded during patient care. These naturally occurring labels present opportunities for broader adoption of supervised approaches in healthcare contexts.

Finally, due to privacy and ethical constraints, we are unable to release the datasets from the two participating hospitals. To partially address this limitation and support reproducibility, we included parallel experiments using the publicly available MIMIC-IV dataset. This allows external researchers to validate key aspects of our methodology.

Ethical Statement

This study was conducted in collaboration with two partner hospitals and was approved by the respective institutional ethics committees (Helsinki Committees). All data used in this research were handled in accordance with relevant ethical guidelines and institutional policies to ensure patient privacy and data protection.

References

Xue B. Abraham J. et al. Alba, C. 2025. The foundational capabilities of large language models in predicting postoperative risks using clinical notes. *npj Digital Medicine*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bryan Perozzi Bahare Fatemi, Jonathan Halcrow. 2024. Talk like a graph: Encoding graphs for large language models. *ICLR*.

Harshavardhan. Battula, Jiacheng. Liu, and Jaideep. Srivastava. 2024. Enhancing in-hospital mortality prediction using multi-representational learning with LLM-generated expert summaries. *arXiv*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. Agarwal. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Shan Chen Che Liu5 Zhongwei Wan Danielle S. Bitterman Fei Wang Kai Shu Canyu Chen, Jian Yu. 2024. ClinicalBench: Can llms beat traditional ml models in clinical prediction? *arXiv*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Praveen. Munjal Prateek. Raha Tathagata. Hayat Nasir. Rajan Ronnie. Al-Mahrooqi Ahmed. Gupta Avani. Umar-Muhammad. Gosal Gurpreet. Kanakiya Bhargav. Chen Charles. Vassilieva Natalia. Ben Amor Boulbaba. Pimentel Marco. Christophe, Clément. Kanithi. 2024. Med42-evaluating fine-tuning strategies for medical LLMs: Full-parameter vs. parameter-efficient approaches. *AAAI*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A Miller, Danielle Bitterman, Matthew Churpek, and Majid Afshar. 2024. When raw data prevails: Are large language model embeddings effective in numerical data representation for medical machine learning applications? *EMNLP*, pages 5414–5428.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

M. Gupta, B. Gallamoza, N. Cutrona, P. Dhakal, R. Poulain, and R. Beheshti. 2022. An extensive data processing pipeline for MIMIC-IV. *Proceedings of Machine Learning Research (PMLR)*, 193:311–325.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453, Singapore. Association for Computational Linguistics.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. TabLLM: Few-shot classification of tabular data with large language models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2022. MIMIC-IV (version 2.0).

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.

Patricia Cabanillas Silva Mohamed Rezk and Fried Michael Dahlweid. 2024. LLMs for clinical risk prediction. *"ArXiv"*, arXiv:2409.10191.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2024. GPT-4: Large language model. `https://chat.openai.com/chat`. Accessed: 2025-05-13.

Yingce Xia Tao Qin Sheng Zhang Hoifung Poon Tie-Yan Liu Renqian Luo, Liai Sun. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23.

Alicia Russell-Gilbert, Alexander Sommers, Andrew Thompson, Logan Cummins, Sudip Mittal, Shahram Rahimi, Maria Seale, Joseph Jaboure, Thomas Arnold, and Joshua Church. 2024. Aad-llm: Adaptive anomaly detection using large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4194–4203.

Abhilash Shankarampeta, Harsh Mahajan, Tushar Kataria, Dan Roth, and Vivek Gupta. 2025. TRANSIENTTABLES: Evaluating LLMs' reasoning on temporally evolving semi-structured tables. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6526–6544, Albuquerque, New Mexico. Association for Computational Linguistics.

David Salinas Frank Hutter Shi Bin Hoo, Samuel Müller. 2025. The tabular foundation model TabPFN outperforms specialized time series forecasting models based on simple features. *"ArXiv"*.

Ofir Ben Shoham. and Nadav Rappoport. 2025. CPLLM: Leveraging LLMs for data representation in clinical prediction. *ArXiv*.

Rishivardhan Krishnamoorthy Avi Patel Gabriel Wardi Joseph C. Ahn Karandeep Singh Eliah Aronoff-Spencer Shamim Nemati Supreeth P. Shashikumar, Sina Mohammadi. 2025. Development and prospective implementation of a large language model based system for early sepsis prediction. *medRxiv*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*.

Zhang X Zhang Y Xie W Wang Y. Wu C, Lin W. 2024. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association : JAMIA,*.

Chen Q. Chen A. Peng C. Hu Y. Lin F. Peng X. Huang J. Zhang J.-Keloth V. Zhou X. He H. Ohno-Machado L. Wu Y. Xu H. Bian J. Xie, Q. 2024. Me-LLaMA: Foundation large language models for medical applications (version 1.0.0). *PhysioNet*.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. 2024. Me LLaMA: Foundation large language models for medical applications. *Preprint*, arXiv:2402.12749.

Hao Xue and Flora D. Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *Preprint*, arXiv:2210.08964.

Daniel McDuf Cynthia Breazeal Hae Won Park Yubin Kim, Xuhai Xu. 2024. Health-LLM: Large language models for health prediction via wearable sensor data. *Proceedings of Machine Learning Research (PMLR)*, arXiv:2401.06866.

A Appendix: Datasets

The MIMIC-IV dataset is made available under the PhysioNet Credentialed Health Data License (CHDL), which requires data access approval and adherence to specific ethical guidelines. Access to the dataset can be requested via PhysioNet [#]. We ensured that our use of MIMIC-IV aligns with their intended purpose as specified by their respective licenses. Specifically, we used it only for research-oriented use.

In-addition, We provide additional information about the data, including label distributions for each dataset (Table 4) and a snapshot tabular data example with synthetic data (Table 8).

Feature Set We utilized a comprehensive set of clinical features derived from structured EHR data, encompassing demographic information, medication administrations, procedural interventions, vital signs, and laboratory measurements. The selected features include age, gender, insurance type, and specific clinical markers such as pH, pCO_2, pO_2, oxygen saturation, lactate, hemoglobin, white blood cells, and platelet count.

Dataset	**Split**	**# Instances**		**Positive**
		0	**1**	**Ratio**
TASMC	Train (all)	112,021	76,134	0.40
	Train	26,898	18,261	0.40
	Validation	14,191	6,051	0.30
	Test	12,653	6,495	0.34
ShebaMC	Train (all)	152,765	95,763	0.38
	Train	30,553	19,153	0.38
	Validation	20,982	9,971	0.32
	Test	23,203	12,869	0.35
MIMIC-IV	Train (all)	490,407	317, 776	0.39
	Train	24,520	15,889	0.39
	Validation	29,711	20,831	0.41
	Test	34,781	21,219	0.38

Table 4: Label Distributions.

Additionally, we included a range of laboratory values such as creatinine, glucose, albumin, C-reactive protein, and alkaline phosphatase, as well as interventions like vasopressor and antibiotic administration, endotracheal intubation, and urinary filtration procedures. The complete list of features is detailed in Table 7.

B Appendix: Training Setup

Training was performed on a server equipped with an NVIDIA A10G GPU (96GB VRAM). A single experiment (seed), including training and inference, took approximately 4 days on a single 24GB GPU. In total, we utilized around 9000 hours of 96GB GPU. To improve memory efficiency and training speed, we applied Low-Rank Adaptation (LoRA) (Hu et al., 2022) with 4-bit quantization for all LLM experiments. The fine-tuning process was conducted using a dual approach, where the model was trained on both the context and the response. The number of tokens varied across different verbalization configurations, ranging from approximately 1,000 tokens for the *Forward-Fill* aggregation method to 3,000 tokens for the *per-event* method.

To determine the best-performing setup, we conducted a grid search over our core generative configuration using the validation set. The optimal hyperparameters were then fixed and used in all subsequent experiments. Specifically, we tuned the learning rate, batch size, LoRA rank, LoRA alpha, and weight decay. Final values for these hyperparameters are shown in Figure 3.

A grid search was also performed to optimize the baseline models. For Logistic Regression, we tuned regularization strength, solver type, and convergence parameters. For LSTM, we optimized network architecture (depth and layer sizes), sequence

[#] https://physionet.org/content/mimiciv/2.0/.

```
num_train_epochs: 5
learning_rate: 2e − 04
gradient_accumulation_steps: 16
per_device_train_batch_size: 2
per_device_eval_batch_size: 2
max_seq_length: 3000
lora_rank: 64
lora_alpha: 32
weight_decay: 0.01
warmup_ratio: 0.06
max_grad_norm: 0.3
```

Figure 3: Final parameters for Llama-3-8B-Instruct fine-tuning process.

length, learning rate, batch size, and regularization parameters (L2, dropout). For XGBoost, we tuned tree depth, learning rate, and regularization parameters. Final selected hyperparameters are provided in the following figure 4.

Logistic Regression
penalty: L2
C: 0.001
solver: liblinear
max_iter: 1000

LSTM
layers: [16,16,16,1]
history: 5 steps
learning_rate: 10^{-3}
batch_size: 512
L2: 0.03
dropout: 0.2

XGBoost
learning_rate: 0.05
n_estimators: 100
max_depth: 5
subsample: 1.0
reg_lambda: 10

Figure 4: Final hyperparameters for baseline models selected via grid search.

C Appendix: Me-LLaMA Baseline

To validate the effectiveness of our fine-tuning process, we use `Me-LLaMA` as a baseline, as it has demonstrated notable improvements over other open-source models, such as `GPT-4` and `LLaMA`, through continued domain-specific training on clinical datasets, including MIMIC-IV. Optimized for biomedical and clinical text analysis, `Me-LLaMA` leverages extensive pre-training to enhance medical reasoning and diagnostic capabilities, making it a robust reference for evaluating our fine-tuning approach in healthcare contexts.

Establishing a reliable baseline involves two key steps: (1) clearly defining the task for the model, as it is employed in a zero-shot setting, and (2) clearly specifying the response format to ensure that the binary label is among the top two predicted tokens for accurate evaluation. After several iterations, we identified the prompt in Figure 10 as the optimal formulation to align with our objectives. Then, we apply the same inference process outlined in Figure 1.

As a further step, we evaluated Me-Llama using the verbalization variants proposed in our paper. Although we observe an improvement when applying per-event aggregation with exact timestamp annotations and a history length of six, the results remain far below the performance achieved by the fine-tuned XGBoost or generative model. Specifically, Me-Llama with this configuration achieved an AUC of 68.1% on MIMIC-IV and 66.48% on TASMC.

D Appendix: Narrative Style

The objective of the narrative strategy is to leverage the inherent strengths of LLMs in processing textual data while maintaining the aggregation frameworks developed for structured data. The aim is to transform structured data into a more descriptive, narrative form, allowing for a nuanced representation of indicator or event trajectories.

Per Indicator Narrative Approach. For the *per-indicator* aggregation method, the objective is to present the trajectory of each indicator in a time series format. The narrative is structured deterministically, focusing solely on the start and end times of the observed period. This approach preserves the chronological order while emphasizing key changes.

Per Event Narrative Approach. In contrast, the *per-event* narrative strategy aims to provide a more descriptive account of each indicator. The approach introduces two risk thresholds to classify the risk level for each indicator. Using these thresholds, GPT-o1 generates a descriptive sentence for each indicator, outlining its observed value and the associated risk level. These sentences serve as templates and are applied deterministically to all instances. This narrative can vary in complexity, ranging from a straightforward low-risk statement to a more detailed account that may include potential diagnoses or critical observations. See Figure 9 for an illustrative example.

Table 5 presents the results for the two narrative prompting approaches.

Experiment	AUC-ROC	
	TASMC	MIMIC-IV
	Technical	
Per Indicator *best*	77.78±0.13	84.93±0.12
Per Event *best*	77.94±0.31	89.68±0.18
	Descriptive	
Per Indicator *best*	77.51±0.13	84.47±0.41
Per Event *best*	77.44±0.02	90.40±0.04

Table 5: Comparison between the best-performing *technical* narrative and the *descriptive* narrative.

E Appendix: Analyzing the Impact of MIMIC's Composite Label on the Performance of the Model

We found that both *per-indicator* and *per-event* aggregation methods outperformed the *forward-fill* baseline across tasks, with *per-event* consistently achieving better results in the MIMIC-IV task—showing an approximate 5% improvement across all prompt configurations.

To better understand the impact of *per-event*, we evaluated its performance on individual labels—heart failure and mortality—while keeping all negative instances intact and isolating positive instances for each label (see Figure 6).

In the composite MIMIC-IV task, which includes 30-day heart failure, *Per Event* aggregation method remained the top performer. However, when focusing on individual outcomes, *Per Indicator* was competitive for mortality prediction, while the performance gap widened for 30-day heart failure prediction.

These results suggest that although both aggregation methods improve the mortality outcome prediction, *Per Event* provides a clear advantage for heart failure.

F Appendix: Single Snapshot per Admission Evaluation

In this section, we evaluate the models using a single time point (snapshot) per admission to assess robustness at the admission level.

We consider two settings as presented in Figure 7: the first selects a snapshot taken 84 hours after admission, while the second uses the last available time-point for each admission (for positive cases we use the last positive snapshot).

Experiment	AUC-ROC	
	In-Hospital Mortality	Heart Failure
Forward-fill (XGBoost)	78.25	80.08
Forward-fill (Llama)	79.64	80.45
Per Indicator *best*	**80.57**	81.35
Per Event *best*	**81.01**	**87.91**

Table 6: Evaluating the Mimic-IV composite task on specific outcomes separately, we observe improvements from both *per-indicator* and *per-event* methods for mortality prediction. However, a larger performance gap emerges in heart failure prediction, highlighting the advantage of the *per-event* approach.

The findings indicate consistency with our primary evaluation setting across both datasets. Furthermore, we observed that the *per-event* method outperforms other aggregation strategies on the MIMIC-IV dataset, particularly in early detection scenarios, achieving a 13% performance gain at the 84-hour post-admission setting.

Additionally, in Figure 5, we present results for the TASMC dataset using alternative timestamps (54, 108 and 156 hours). The results further confirm the robustness of the *per-event* method, with a consistent improvement, highlighting the method's effectiveness regardless of the specific timestamp.

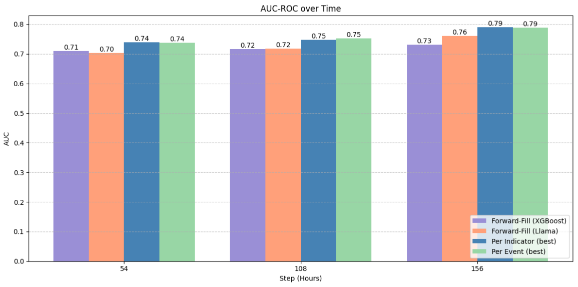


Figure 5: Single snapshot evaluation per admission at 54, 108, and 156 hours, selected as 6-hour interval multiples in TASMC.

G Appendix: Verbalizer Examples

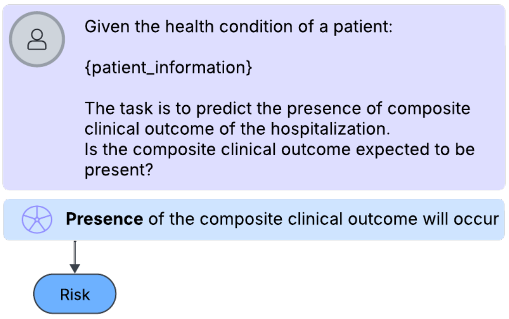
We provide examples of the verbalizer output based on our configurations (Figure 8), along with the prompt template (Figure 6) and a fully populated prompt example for the descriptive approach (Figure 9). Due to ethical considerations, all examples use synthetic values.

Feature Category	**Features**
Demographics	Age, Gender, Insurance
Medications	Vasopressors, Antibiotics, Diuretics, Analgesics
Procedures	Endotracheal Airway Insertion, Infusion Device Insertion, Nutritional Substance Introduction, Urinary Filtration, Coronary Artery Fluoroscopy
Vitals and Labs	pH, pCO_2, pO_2, Oxygen Saturation, Lactate, Hemoglobin, Hematocrit, White Blood Cells, Platelet Count, Eosinophils, Creatinine, Urea Nitrogen, Sodium, Potassium, Bicarbonate, Anion Gap, Glucose, Albumin, Calcium, Alkaline Phosphatase, C-Reactive Protein, D-Dimer

Table 7: Selected Clinical Indicators from MIMIC-IV Used in Model Training

TimeFromHosp	**Lactate**	**pH**	**Hemoglobin**	**WBC**	**Creatinine**	**Bicarbonate**	**Calcium**	...
	(mmol/L)		(g/dL)	(10^3/μL)	(mg/dL)	(mEq/L)	(mg/dL)	
0 days 04:00	1.4	7.36	9.8	13.3	0.9	24	8.1	...
0 days 12:00			10.1	12.9				...
1 days 00:00	1.2		10.4	16.1	1.1		9.6	...
1 days 12:00		7.35				22		...
2 days 00:00				14.2	1.3		8.7	...
2 days 08:00	0.9	7.38	10.6	14.8		20	9.1	...
2 days 16:00				12.4				...
3 days 04:00			10.3	10.8		23	9.7	...
3 days 12:00	1.1				1.0			...
4 days 00:00	1.3		10.0			25		...

Table 8: A synthetic example representing a *snapshot*, instance, of patient data over 96 hours post-admission, illustrating the sparsity and varying update frequencies across clinical indicators. Empty cells are missing values



Given the health condition of a patient:

{patient_information}

The task is to predict the presence of composite clinical outcome of the hospitalization.
Is the composite clinical outcome expected to be present?

Presence of the composite clinical outcome will occur

Figure 6: Predicting a patient's risk of a clinical outcome using an LLM. Given a prompt containing the patient's journey data from hospitalization, our fine-tuned model generates a response with a binary phrase classification. The predicted probability of the first token determines the patient's risk for a clinical outcome.

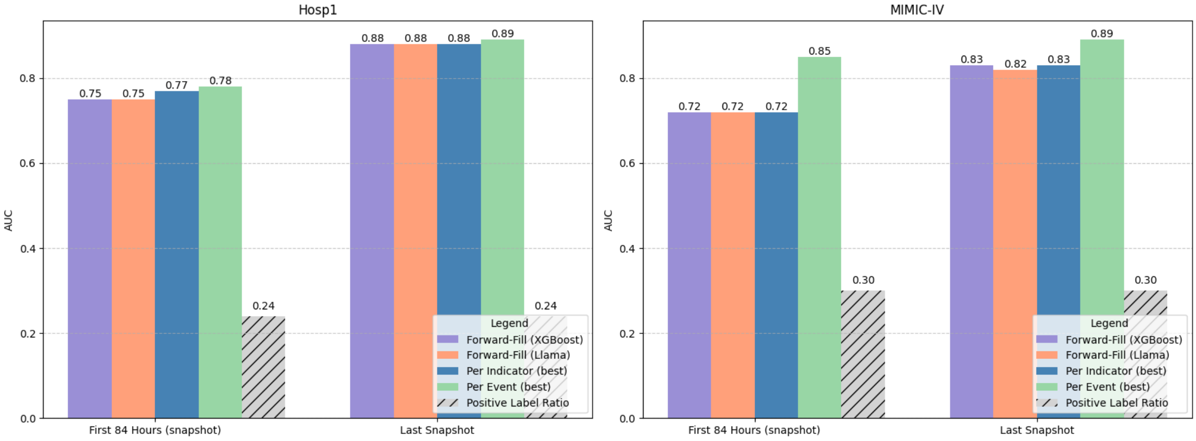
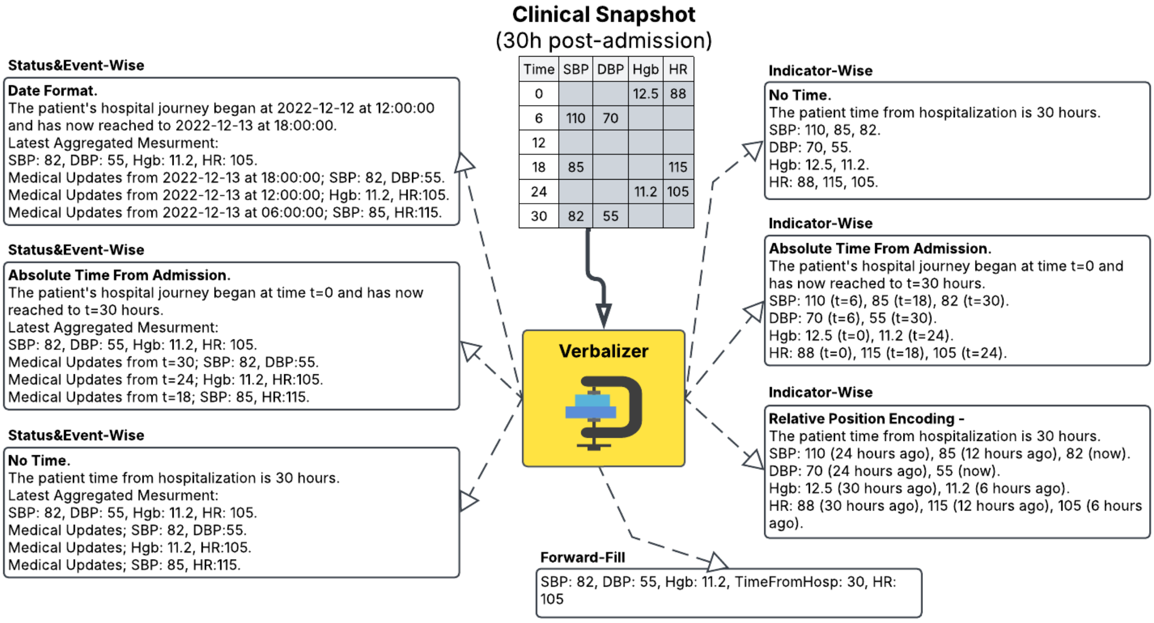


Figure 7: Evaluation of a single snapshot per admission. (1) First Snapshot (84 hours): We select the snapshot at 84 hours or the last available snapshot for shorter admissions. (2) Last Snapshot: We define the last positive snapshot (prior to treatment initiation). If no positive snapshot is present, we use the final snapshot.



Clinical Snapshot
(30h post-admission)

Time	SBP	DBP	Hgb	HR
0			12.5	88
6	110	70		
12				
18	85			115
24			11.2	105
30	82	55		

Status&Event-Wise

Date Format.
The patient's hospital journey began at 2022-12-12 at 12:00:00 and has now reached to 2022-12-13 at 18:00:00.
Latest Aggregated Mesurment:
SBP: 82, DBP: 55, Hgb: 11.2, HR: 105.
Medical Updates from 2022-12-13 at 18:00:00; SBP: 82, DBP:55.
Medical Updates from 2022-12-13 at 12:00:00; Hgb: 11.2, HR:105.
Medical Updates from 2022-12-13 at 06:00:00; SBP: 85, HR:115.

Status&Event-Wise

Absolute Time From Admission.
The patient's hospital journey began at time t=0 and has now reached to t=30 hours.
Latest Aggregated Mesurment:
SBP: 82, DBP: 55, Hgb: 11.2, HR: 105.
Medical Updates from t=30; SBP: 82, DBP:55.
Medical Updates from t=24; Hgb: 11.2, HR:105.
Medical Updates from t=18; SBP: 85, HR:115.

Status&Event-Wise

No Time.
The patient time from hospitalization is 30 hours.
Latest Aggregated Mesurment:
SBP: 82, DBP: 55, Hgb: 11.2, HR: 105.
Medical Updates; SBP: 82, DBP:55.
Medical Updates; Hgb: 11.2, HR:105.
Medical Updates; SBP: 85, HR:115.

Indicator-Wise

No Time.
The patient time from hospitalization is 30 hours.
SBP: 110, 85, 82.
DBP: 70, 55.
Hgb: 12.5, 11.2.
HR: 88, 115, 105.

Indicator-Wise

Absolute Time From Admission.
The patient's hospital journey began at time t=0 and has now reached to t=30 hours.
SBP: 110 (t=6), 85 (t=18), 82 (t=30).
DBP: 70 (t=6), 55 (t=30).
Hgb: 12.5 (t=0), 11.2 (t=24).
HR: 88 (t=0), 115 (t=18), 105 (t=24).

Indicator-Wise

Relative Position Encoding -
The patient time from hospitalization is 30 hours.
SBP: 110 (24 hours ago), 85 (12 hours ago), 82 (now).
DBP: 70 (24 hours ago), 55 (now).
Hgb: 12.5 (30 hours ago), 11.2 (6 hours ago).
HR: 88 (30 hours ago), 115 (12 hours ago), 105 (6 hours ago).

Forward-Fill

SBP: 82, DBP: 55, Hgb: 11.2, TimeFromHosp: 30, HR: 105

Figure 8: Examples of the verbalizer output based on our configurations.

Per Event Narrative Strategy Example

The patient's hospital journey began on 2022-12-02 at 12:00:00 and has now reached 2022-12-06 at 16:00:00.
Latest Aggregated Measurement:
Lactate is within or near normal range at 1.2 $mmol/L$.
Blood pH is normal at 7.25.
pO2 is above 100 mmHg at 108.0, possibly due to high oxygen supplementation.
Oxygen saturation is adequate or high at 98.0%.
Hemoglobin is low at 9.0 g/dL, suggesting anemia.
Hematocrit is low at 26.2%, indicating possible anemia.
WBC is elevated at 16.8 $x10^3/uL$, suggesting infection or inflammation.
Platelet count is low at 88.0 K/uL (thrombocytopenia).
Creatinine is elevated at 1.3 mg/dL, suggesting renal impairment.
BUN is within normal range at 11.0 mg/dL.
Anion gap is normal at 12.0 mEq/L.
Sodium is within normal range at 135.0 mEq/L.
Potassium is within normal range at 4.3 mEq/L.
Glucose is within normal range at 130.0 mg/dL.

Medical updates from 2022-12-06 at 12:00:00: Sodium is low at 133.0 mEq/L (hyponatremia). Potassium is within normal range at 4.1 mEq/L.

Medical updates from 2022-12-06 at 00:00:00: Blood pH is normal at 7.25. pCO2 is within normal range at 36.0 mmHg. pO2 is above 100 mmHg at 106.5, possibly due to high oxygen supplementation. Oxygen saturation is adequate or high at 98.0%.

....

Figure 9: Illustration of Narrative Strategy for *per-event* Aggregation.

For MIMIC-IV task

Your task is to predict whether a composite clinical outcome will occur based on the patient's health condition.
A composite clinical outcome is considered present if any of the following clinical outcomes are observed:
1 - In hospital mortality
2 - Heart failure in 30 days
3 - Length of stay > 15 days
Patient's health condition: text

Indicate the composite clinical outcome: 0 - absent, 1 - presence.
The composite clinical outcome is expected to be

For TASMC and ShebaMC task

Your task is to predict whether a composite clinical outcome will occur based on the patient's health condition.
A composite clinical outcome is considered present if any of the following clinical outcomes are observed:
1 - In hospital mortality
2 - Transfer to ICU/SDU
3 - Administration of inotropic medications
4 - Length of stay > 15 days

Indicate the composite clinical outcome: 0 - absent, 1 - presence.
The composite clinical outcome is expected to be

Figure 10: Me-LLaMA prompt template for zero-shot inference.