MIMIC-EXT-TS: A Discrete Clinical Time-Series Dataset for ICU Patients

Anonymous ACL submission

Abstract

Early detection and timely treatment are critical in medicine. For example, surgical excision of skin lesions can cure early-stage skin cancer, but once metastasis occurs, even the most advanced therapies often fail. In this work, we introduce MIMIC-EXT-TS, a large-scale dataset with an over 11 million clinical event, timestamp pairs from over 267k free-text discharge summaries. It is the first discrete time series clinical events dataset of ICU patients. It significantly reduces the storage time and presents important clinical events and related timestamp in structured way. To achieve the dataset, we propose an end-to-end RAG model with an LLM agent for temporal reasoning. The integrity check confirms over 94% of events are can be traced back to source clinical note with mean token overlap around 91%. All important events, such as diagnosis ICD codes are captured. We further validate the effectiveness of our dataset in downstream tasks, and fine-tuning LLMs, such as Qwen and MedGemma, for tasks on MedMCQA, MMLU, and PubMedQA dataset.

1 Introduction

The timing of medical symptoms is with critical importance for clinical risk prevention. Consider a solid tumour such as breast cancer, when identified at an *early, localized* stage, curative surgical resection is often sufficient and relatively low in morbidity. Once the tumour has metastasized via hematogenous spread to bone or other organs, management shifts to multi-agent systemic therapy such as chemotherapy, endocrine or targeted agents. A cure is much less likely. To develop the machine learning model that helps predict the potential timeline of the risk, the dataset with clinical events and related timeline is recommended.

Though structured electronic healthcare records have temporal information for events, such as diagnosis ICD codes and lab test, the important temporal information such as the timestamp of the first observation of the symptom is not available. The tabular data are also limited to the predefined ICD codes. The clinical notes, such as discharge summaries, record more detailed information which are important for treatment.

There are many efforts to extract the clinical events and timestamps. For example, the 2012 i2b2 Temporal Relations Challenge [28] released the dataset with clinically significant events, temporal expressions and temporal relations, such as "before", "after". However, the temporal relations are limited to the two clinical events, such as "stress dose steroids BEFORE his surgery". The challenge considers temporal expressions which refer to dates, or frequencies phrases in the clinical text. However, direct temporal expressions for events are usual missing in the note. The physicians usually use their experience and expertise to infer the timestamp of event which takes a long time. Hence, based on the free-text clinical notes or current structured electronic health record, it is challenging to find the causal inference or cofounder of disease based on local temporal expression. In this work, we solve the the challenge by introducing a discrete time-series representation for the history of patients, that is, a sequence of clinical events and related timestamps. We choose a pivot event such as admission to the hospital or the main diagnosis, the timestamp of all other events are converged to negative (history) and positive (future) in hours. It is easy to find the history and future treatment plan based on the temporal information. Table 1 shows a synthetic example of our temporal representation of discharge summary from MIMIC-IV.

Formally, we represent the trajectory of patient with a sequence of time-event pairs $\{T_i, E_i\}_{i\geq 1}$. Event E_i is a text span that can be anything related to patients, such as symptoms, lab tests, and treatments. The time T_i is the timestamp that the event E_i occurred. To track the correlation

Table 1: The synthetic structured clinical events and timestamps of a patient. The pivot events are record in the last two rows, with timestamp 0. The first row indicates that 72 hours prior to the pivot event, the patient underwent high-grade fever.

Timestamp	Clinical events
−72.0	High-grade fever (38.9 °C)
−72.0	Truncal maculopapular rash
−72.0	Acute pyelonephritis
−72.0	Aspiration pneumonia
−72.0	Hypoxia (SpO_2 86% on room air)
−72.0	Metabolic acidosis (pH 7.25)
−72.0	Hypotension (MAP 59 mmHg)
−72.0	Profuse diarrhea (>6 stools/day)
0	DNI order placed
0	Code blue (PEA arrest)

between important clinical events, we propose a global timestamp representation by defining major clinical events, such as hospitalization, as pivot event. The events before a pivot event are with negative timestamp values, and the future events, such as "follow-up visits" in 3 months, are with positive timestamp values. The timestamps are converted to the same unit, hour. Our released dataset MIMIC-EXT-TE, consists of 11,784,557 clinical time-event pairs from 267,268 visits. Here is a summary of our contributions:

1. A discrete time series dataset with a comprehensive time-event representations of each in-hospital stay for patients from MIMIC-IV [12, 11], which is a publicly available database sourced from the electronic health record of the Beth Israel Deaconess Medical Center. Our dataset consists of 11,784,557 time-event pairs.

2. To create this dataset, we build an end-to-end framework with an LLM agent to automatically convert free-text note to a sequence of event and related timestamps pairs. That is, the framework first retrieves clinical events by Retrieval-Augmented Generation (RAG) with a brief summary of the note as query. Then a Large Language Model (LLM) agent extracts the temporal information of events based on chain-of-thought reasoning on source note.

3. By integrity checking, over 94% of events in our dataset are aligned with source note with mean token overlap around 91%. All the diagnosis ICD in MIMIC-IV are captured by our dataset. Human evaluation on 4,365 events confirm over 94% matched. The mean average of temporal difference on 2248 unique chart events with recorded time is 32 hours. It shows that our dataset is aligned with the original clinical notes and our pipeline is robust to hallucinations.

4. We compare the performance of our dataset and original clinical note on important downstream tasks, such as in-hospital mortality and 30-day readmission, our dataset achieves significant improvement. It demonstrates the high quality of our dataset with most important information.

5. To validate our contributions to LLM community, we fine-tune, a general LLM model Qwen3 [29] and a medical LLM, MedGemma [27] on our dataset. Then we evaluate their performance on datasets such as MedMCQA [23], MMLU [8] and PubMedQA [10]. The fine-tuned Qwen3-4B achieves 18% boost in accuracy on MedMCQA. The fine-tuned MedGemma does not guarantee performance improvement. The LLMs fine-tuned on both PubMed and our dataset achieves the best performance in most cases. It demonstrates the contributions of our dataset on patient related clinical/medical question answering.

2 Main Result

We propose a discrete temporal presentation of free-text clinical note of ICU patient's in-hospital stay from MIMIC-IV dataset.Our dataset converts the 267,268 free-text discharge summaries into a sequence of 11 million event-time pairs. To achieve this goal, we propose an end-to-end retrieval framework with an LLM agent for temporal reasoning.

We chose MIMIC-IV because it focuses on ICU patients, where timely treatment is critical to patient survival. The treatment decisions and clinical trajectories of ICU patients are valuable for improving auto healthcare, including applications such as developing medical LLM agents. Discharge summaries provide some of the most comprehensive documentation of a patient's in-hospital course and therefore serve as our primary data source. However, discharge summaries are often lengthy and highly unstructured, making automated informa-

tion extraction challenging. In particular, temporal details about clinical events are rarely stated explicitly and frequently must be inferred from context. As shown in Figure 1, the highlighted spans mark the temporal expressions used in our timeline construction.

> The patient is a 47-year-old female with a documented history of alcohol use disorder who presented to the emergency department with a **5 day** history of worsening abdominal pain. She reports a recent period of increased alcohol consumption lasting **six weeks**, during which she drank an estimated 12 to 15 standard drinks **daily**. This period of increased intake ended approximately **three weeks ago**.

Figure 1: A synthetic clinical note.

Hence, our discrete time series dataset presents a summary of important events. It is convenient for physicians to review the trajectories of patient. The structured dataset is easier to be used in other machine learning based healthcare applications. To extract the important clinical events and temporal information, we propose to retrieve events by RAG and utilize an LLM agent for temporal information reasoning. The illustration of the end-to-end framework is shown as Figure 2. The query is Brief Hospital Courses (BHC), which are succinct summaries of an entire hospital encounter, written by senior clinicians responsible for the overall care of a patient [26], hence, the name entities with high semantic correlation with BHCs are important clinical events. The admission is served as the pivot event. With the pivot events available, we use LLMs to read the clinical note and find timestamps of important events based on the pivot event.

To evaluate the quality of the dataset, we apply integrity check to find the alignment of our dataset and data source. We use extract match in the event level and admission level to find over 90% of events can be traced back to the original notes. The high alignment with temporal information with chartevent table from MIMIC-IV shows our LLM agent does a great job in temporal reasoning. It validates the high quality of our dataset. It also demonstrates the significant advantage of our pipeline with an LLM based temporal reasoning agent. Then we test the performance on the downstream tasks, such as in-hospital mortality and 30-day readmission, with our dataset and clinical notes only. Our dataset achieves 88.1% accuracy in mortality prediction compared with 82.1% with clinical note. The clinical note reaches 61.8% in 30-day readmission prediction while ours is 59.8%. It shows that our dataset captures most important events and related temporal information. To demonstrate the clinical knowledge of our dataset, we fine-tuned Large Language Models Qwen and MedGemma on our dataset and test on medical datasets. We use Inverse Cloze Task [16] to create a subset for LLM fine-tuning. That is, we first select discharge summary with at least at least 10 clinical event-time pairs, and randomly choose one event-time pair as "answer" and others as context. The process is repeated 10 time without duplication of the "answer". It leads to a dataset with 2 million event-time pairs. We fine-tune Qwen3-4B with the subset, and achieve 18% superiority in accuracy compared with vanilla Qwen (71.84% vs. 53.81%) on medical examinations dataset, MedMCQA. MedGemma achieves consistent gains across almost all specialties (e.g., Forensic Medicine 81.9% vs 94.1%, Psychiatry 81.6% vs 94.4%, Skin 87.8% vs 97.5%). MMLU dataset is not patient specific, however, fine-tune on both MMLU and our dataset does not hurt the performance. PubMedQA and our dataset helps achieve the best performance.

3 Related works

There are many medical datasets with temporal information, but most are tabular rather than narrative. For example, the NIH RECOVER program for Long COVID provides visit-level timelines and symptom status per follow-up [30]. While valuable for cohort analyses, such tables typically mark whether a symptom was present at each visit but rarely encode when the symptom first began or the fine-grained temporal relations across events. In contrast, free-text clinical records (e.g., discharge summaries) often contain richer temporal detail (onset, duration, recency, ordering), but this information is entangled in narrative form and requires expert inference to recover. We build on MIMIC-IV-Note [11], whose de-identified discharge summaries contain precisely these temporal cues, our contribution is to structure them into event–time pairs on a unified relative timeline.

Temporal annotation corpora for clinical application. [18] presented a diagnostic video-text dataset for the evaluation of temporal concept understand-

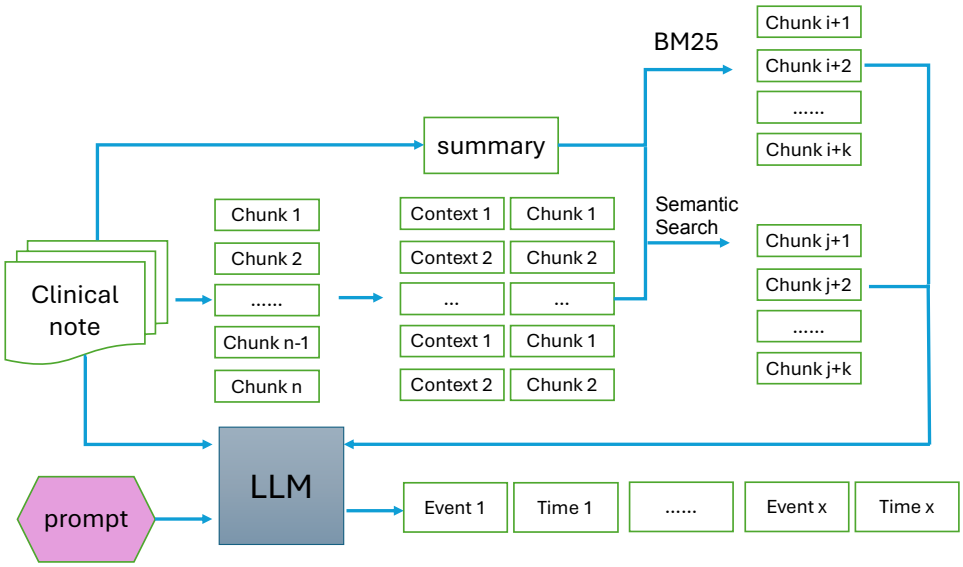

Figure 2: The pipeline of an end-to-end annotation framework with an LLM agent.

ing. The 2012 i2b2 challenge focused on temporal relations in clinical narratives, releasing discharge summaries annotated for events, temporal expressions, and relations (e.g., before/after/overlap) [28]. Beyond i2b2, the THYME project produced oncology notes annotated with TimeML-style events and temporal links, enabling the Clinical TempEval shared tasks [9, 3]. These corpora emphasize calendar-style temporal expressions (dates, times, durations, frequencies) and local pairwise relations. By contrast, our dataset normalizes all events onto a global, relative timeline (hours) centered on a clinical pivot (e.g., admission), which makes patient trajectories and cross-event temporal ordering directly comparable.

Event and temporal information extraction. Clinical event extraction has been tackled via classic and neural IE pipelines and joint models. There are many related works in this direction [20, 38]. Biomedical NER has advanced coverage of entities such as diseases, chemicals, and genes (e.g., CHEMDNER and follow-ups), yet many works remain bounded by fixed type sets, limiting their utility for broad clinical course modeling [39, 25]. Temporal IE lines of work include temporal expression tagging (e.g., HeidelTime-clinical), temporal relation extraction [21, 31], and timeline construction from narrative text [17]. Outside the clinical domain, temporal commonsense work studies typical event durations and orderings [40, 5, 6, 19]. Our approach complements these by (i) anchoring disparate events to a single signed axis (past–future relative to admission) and (ii) emphasizing patient-level trajectories rather than only local pairwise links.

Prompt strategy with expert exemplars. Prompting has progressed from manual templates and few-shot in-context learning [4] to reasoning-oriented prompts such as chain-of-thought (CoT) and zero-shot CoT variants [35, 15], with stability improvements from self-consistency [33]. Human input plays a central role in both *prompt design* and *model alignment*: community toolkits (e.g., PromptSource) and instruction collections aggregate large sets of *human-written* prompt templates and exemplars [2, 34]. At the model-level, human preference feedback (RLHF/DPO) aligns responses to practitioner intent [22, 24]. In this work, we adopt a lightweight, domain-specific variant: a small number of *expert-written exemplars* that demonstrate the desired extraction format and evaluation criteria. This human-in-the-loop prompt design complements CoT-style reasoning when useful, while remaining simple enough for routine clinical annotation workflows.

4 Our method

LLM agent has been widely used for annotation, however, the outputs tend to be random and unreliable. To reduce uncertainty, we propose to retrieve events from original clinical notes, then the LLM is applied to estimate the related timestamp with our designed prompt strategy. In summary, there are two primary steps in constructing our dataset: the clinical events retrieval and temporal information reasoning. For the first step, we use RAG. For the

Original chunk:
"clear to auscultation bilaterally, no".

Our contextualized chunk:
"Neck: cervical lymphadenopathy; supple, JVP not elevated. Lungs: ***clear to auscultation bilaterally, no*** *wheezes, rales, rhonchi. CV: regular rate and rhythm.*

Figure 3: An example of the original and our contextualized chunk demonstrating clinical findings related to lung examination.

second step, we use Llama-3.1-8B powered CoT reasoning[7].

4.1 Clinical Events Retrieval

The retrieval pool is the clinical note. The query is brief hospital course (BHC), which is a section of the clinical note, and acts as a summary of the whole note. Formally, each summary D_j is paired with its BHC query Q_j. The algorithm first breaks the clinical note D_j into a set of chunks C_j with at least $m = 5$ tokens. Then it performs lexical retrieval with BM25 to capture top $k = 100$ matches, that is R_j^{BM25}. Then semantic dense retrieval with a sentence encoder (BGE-Large-en) [37] is applied to recover paraphrastic matches. The encoder converts the query and chunks into 1024-dimensional vectors, and their similarity score is computed. The paraphrastic matches are those with higher similarity score R_j^{emb}. Then we perform re-ranking, which is based on a standard late-fusion scheme to combine BM25 and dense cosine similarities [13, 14], computed as

$$\text{score}(c) \;=\; \lambda{\cdot}\text{BM25}(Q_j, c) + (1{-}\lambda){\cdot}s(c), \quad (1)$$

where $s(c) = \cos\big(\mathcal{E}(Q_j), \mathcal{E}(c)\big)$ and $\lambda \in [0, 1]$. Finally, we return the ranked set

$$S_j \;=\; \text{SortByScore}(U_j), \quad (2)$$

in descending order of $\text{score}(\cdot)$. The union, after deduplication and re-ranking, yields a compact set of 100 candidate chunks per summary. The retrieval process is shown in Algorithm 1.

In practice, we retrieve contextualized chunks rather than isolated spans. For each chunk c, we prepend the previous 10 tokens and append the next 10 tokens, forming $\tilde{c}$. During retrieval, correlations with the query Q (BM25 and dense similarity) are computed on $\tilde{c}$. Although context windows are common in retrieval, they are particularly important in clinical notes, where key findings are distributed across adjacent sentences and sections. Local context often resolves abbreviations, clarifies negations, and links findings across systems (e.g., pulmonary vs. cardiac). As shown in Figure 3, the contextualized chunk yields a more complete examination narrative, covering the neck/JVP, lungs, and cardiovascular systems, thereby supporting the exclusion of alternative etiologies for respiratory symptoms (e.g., heart failure, systemic infection). Without these flanking tokens, the same chunk may appear ambiguous or falsely positive.

Algorithm 1 BHC-Guided Clinical Events Retrieval

Require: Corpus $\mathcal{D} = \{(D_j, Q_j)\}_{j=1}^{N}$; BM25; $\mathcal{E}$ = BGE-Large-en; m=5, k=100, τ=0.75
Ensure: Sets $\{S_j\}_{j=1}^{N}$

for $j \leftarrow 1$ to N **do**
  $C_j \leftarrow \text{Chunk}(D_j,\ \text{min\_tokens} = m)$
  $R_j^{\text{BM25}} \leftarrow \text{TopK}(\text{BM25}(Q_j, C_j), k)$
  for all $c \in C_j$ **do**
    $s(c) \leftarrow \cos(\mathcal{E}(Q_j), \mathcal{E}(c))$
  end for
  $R_j^{\text{emb}} \leftarrow \{\, c \in C_j \mid s(c) \geq \tau \,\}$
  $U_j \leftarrow R_j^{\text{BM25}} \cup R_j^{\text{emb}}$
  Deduplicate U_j keeping the instance with the larger normalized score
  $\text{score}(c) \leftarrow \lambda\,\text{BM25}(Q_j, c) + (1-\lambda)\,s(c)$ for all $c \in U_j$
  $S_j \leftarrow \text{SortByScore}(U_j)$
end for
return $\{S_j\}_{j=1}^{N}$

4.2 Temporal Information Reasoning

Given a discharge summary D_j, and the retrieved contextual chunks S_j, we prompt Llama-3.1-8B to (i) identify clinical events present in the chunks and (ii) assign a relative timestamp (hours) to each event with respect to the pivot event (admission). Events are free-text spans (symptoms, diagnoses, treatments, labs, procedures) that can be temporally located. Here is our timestamp annotation guidance for temporal annotation.

1. The admission event is always assigned a timestamp of 0.

Table 2: Temporal annotation example provided in prompt.

Chunks	Output (event | timestamp)
admitted to the hospital with a 5-day history of fever	`0 | admitted to the hospital`
rash. Four weeks ago, she	`-120 | fever`
was diagnosed with acne	`-120 | rash`
minocycline. This patient	`-672 | acne`
was diagnosed	`-672 | minocycline`
with DRESS syndrome	`0 | DRESS syndrome`

Original clinical note: An 23-year old female was admitted to the hospital with a 5-day history of fever and rash. Four weeks ago, she was diagnosed with acne and received the treatment with minocycline. This patient was diagnosed with DRESS syndrome.

Reasoning: (a) Fever/rash are a 5-day history pre-admission, so timestamp for fever and rash is $-5 \times 24 = -120$ h. (b) Acne/minocycline four weeks prior, so timestamp is $-4 \times 7 \times 24 = -672$ h . (c) DRESS timing not explicit, by clinical judgment near admission, its timestamp is 0.

2. Events occurring before admission have negative timestamps, while those occurring after admission have positive timestamps, measured in hours.

3. Estimate the relative timing of the event based on the context of the entire document.

4. Use explicit or inferred temporal information.

5. When explicit timing is not available, apply clinical judgment to estimate a reasonable time

We also provide a medical expert annotation process as an example as shown in Table 2.

Post-processing pipeline. LLM outputs are high-recall but noisy (e.g., malformed timestamps, swapped columns, duplicates). We apply a deterministic post-processor to enforce format, normalize time, and de-duplicate events. This produces a consistent event-time series per summary. The post-processing pipeline leads to a dataset with 11,784,557 event-time pairs from 267,268 discharge summaries.

5 Statistics of the Dataset

Table 3 summarizes MIMIC-EXT-TS, a large-scale timeline dataset automatically derived from discharge summaries. The coverage of extracted event-time pairs per summary ranges from 1 to 160 events (mean 44), enabling both short and long clinical trajectories to be modeled. Event spans are generally terse (mean 3 tokens) but admit complex phrases (max 286), reflecting everything from single findings (e.g., fever) to multi-clause interventions and outcomes. Events are anchored to a pivot (e.g., hospital admission) and expressed on a common timeline. Most events occur during admission (52.62%), capturing inpatient diagnostics, treatments, and complications; a substantial portion pre-admission (38.08%) reflects prodromal symptoms, prior diagnoses, and antecedent therapies; a smaller tail post-discharge (9.28%) covers follow-up plans, readmissions, and recovery milestones. This distribution supports tasks that depend on when information is observed, not just what is observed, such as early-warning models, causal sequencing, and outcome attribution.

Table 3: Statistics for our dataset.

Statistic	Value
# Discharge summaries	267,268
# Event–time pairs	11,784,557
Min / Max events per summary	1 / 160
Average events per summary	44
Average tokens per event	3
Max tokens per event	286
Temporal distribution of events	
Before admission (historical)	38.08%
During admission	52.62%
After discharge	9.28%

5.1 Integrity check

To assess the consistency and integrity of our extracted events, we compute the exact match between events in our dataset and clinical note. At the corpus level, the exact-match fraction is 77.6% and partial-match fractions are $\geq$94%, with mean token overlap $\approx$91%. At the per-admission level, the median exact-match fraction is 83% and 75% of admissions have exact coverage $>$93%. This shows that nearly all events can be traced back to concrete spans, rather than being free-form LLM hallucinations.

To validate important clinical events are capture, we compare the free-text events to the MIMIC diagnosis table using an Bioclinical BERT [1] based similarity measure. The diagnosis table contained 3,403,042 rows, corresponding to 22,800 unique diagnosis titles. Our dataset has obtained 3,265,485 unique events. For each admission, we computed

Table 4: Comparison results of clinical note and our dataset in downstream tasks, in-hospital mortality and 30-day re-admission.

	Mortality	Readmission
Note-only	81.9	62.3
TS-only	88.1	59.7
Combined	89.9	62.5

the cosine similarity between diagnosis titles and event in the same hospital stay, and declared a match if the similarity based on exceeded a threshold of 0.5. At this similarity threshold, *every* ICD code in the diagnosis table was covered by at least one semantically similar event description. On a per-admission basis, the fraction of diagnoses that were covered by at least one matching event had a mean of 100% with an extremely small standard deviation (8.4×10^{-5}). The minimum coverage fraction across all admissions was 95.65%, and the 25th, 50th, and 75th percentiles were all 100%.

The matched events align reasonably well in time. That is, across 2,248 matched pairs with charttime table from MIMIC, the absolute difference between event time and charttime has a median of 16.1 hours (3.85 to 54.19 hours). The mean difference is 32.76 hours, and nearly all matches fall within 7 days (max 165.47 hours).

5.2 Downstream tasks

We check the multi-task downstream tasks (in-hospital mortality and 30-day readmission) that compare discharge-note-only, event-time-only (our dataset), and combined representations, we show that the extracted event–time representation is both highly informative and complementary to free-text notes. We use Bioclinical BERT [1] for embedding and train a logistic regression model for the tasks. The results are reported in Table 4.

5.3 Human evaluation of factual accuracy

We manually annotated all events extracted from 100 randomly sampled admissions (4,365 events). Clinicians labeled each event as exact matched, partially matched, or not matched by the corresponding discharge summary. The result is 68.3% events are exact matched, 27.0% partially matched, and 4.7% not matched. Thus 95.3% of events were judged supported or partially supported by clinical note, providing a direct human assessment that hallucinations are relatively rare and mostly minor.

6 Experiments

We fine-tuned LLMs with our datasets on medical question answering datasets. We choose a general model Qwen3-4B, and a medical LLM, MedGemma.

6.1 Experiment Design

We generate the subset subTS by inverse close task. That is, we first choose discharge summary with at least 10 time-event pairs, and randomly select one as "answer" and others as context. We repeat the process for 10 times for each discharge summary without duplicate "answer". It leads to the read-to-fine-tuning dataset with 2,528,240 records from 252,824 discharge summaries. Each record is event-timestamp pair.

We use TRL SFTTrainer [32] with HuggingFace Transformers [36]. We use AdamW with cosine decay and warmup ratio 0.03. Unless stated, we train for 3 epoch (step-capped runs are noted separately). We fix random seeds (42) and enable TF32. All experiments on conducted on a server with two GPUs, NVIDIA RTX PRO 6000 and NVIDIA RTX 6000.

6.2 Datasets

MedMCQA [23] is taken from AIIMS & NEET PG entrance exam. The AIIMS PG entrance exam was an entrance examination for postgraduate (PG) medical courses at AIIMS and other institutions. NEET PG (National Eligibility cum Entrance Test for Postgraduate) is a national-level entrance exam in India for medical graduates. The dataset covers 21 medical subjects and over 2,400 healthcare topics. We only consider the single choice questions in this work. There are 138,153 question answer pairs in the released training dataset. Each question has four options, A to D. Since the label of testing dataset is not released. We randomly select 30% from training dataset as our testing and the left as training. Our testing dataset has 56,120 questions in total.

PubMedQA [10] is a biomedical question answering dataset collected from PubMed abstracts. The question is from an existing PubMed article title. The context is the abstract of the article and the answer is Yes, No or maybe. We use the 1,000 human labeled PubMedQA as our dataset. We split the dataset with 70% for training and the left 30% for testing. We repeat the experiment over five randomized splits with random seeds {3407, 1234,

2024, 2025, 777} and report mean result.

For MMLU, we focus on the Clinical Knowledge, College Medicine, and Biology tasks. We fine-tune the model using the default training and validation splits, and report results on the official test split.

6.3 Results

The accuracy of Qwen on MedMCQA different subjects are shown in Table 5 and Table 6 in Appendix. The column dataset means fine-tuning datasets, that is "Base" means the base Qwen model, "subTS" is Qwen fine-tuned on our subset, "MCQA" is Qwen fine-tuned on MCQA training dataset. and "SubTS+MCQA" means Qwen model fine-tuned on both datasets. Averaged across all 21 subjects, the vanilla model attains 56.23% accuracy, while fine-tuning on our subset reaches 73.84%, a gain of +17.61%. The largest absolute improvements appear in *Dental* (+23.07%), *Biochemistry* (+22.77), *Microbiology* (+21.86%), *Medicine* (+20.85%), and *Anatomy* (+20.76%), and achieves over $80\,\%$ in several subjects (e.g., *Biochemistry*, *Psychiatry*, *Skin*).

The results of MedGemma on MedMCQA are shown in Table 7 and Table 8 in Appendix. The results indicate that temporal pretraining alone confers a strong, broadly transferable inductive bias. Overall, the macro-average improves from 77.4% (Base) to 89.5% (MedMCQA), with our subset subTS providing a modest lift to 79.3%, and the combined setting reaching 89.3%.

The results of Qwen and MedGemma on PubMedQA is shown in Table 9. PubMedQA fine-tuning substantially improves performance over the base model, while subTS alone degrades accuracy in this setup; combining subTS with PubMedQA yields the strongest results.

Table 10 summarizes domain-specific multiple-choice accuracy for Qwen and MedGemma under different training setups, evaluated on three MMLU-style subsets: Clinical, Medical Genetics, and College Biology. The training process does not necessarily guarantee performance gains. In our experiments, Qwen fine-tuned on our dataset achieves performance comparable to the base model. This suggests that even for datasets not directly related to patient-level trajectories, fine-tuning on our data does not substantially harm performance.

7 Conclusions

In this work, we introduced a novel discrete time series representation of patient clinical histories, built from event-time pairs extracted from the discharge summaries. By leveraging retrieval-augmented generation for clinical event retrieval and an LLM agent for temporal reasoning, we constructed the dataset containing over 11 million event–time pairs.

The integrity check confirms strong alignment between our constructed dataset and the source clinical notes, highlighting our pipeline's ability to reduce LLM hallucination risk. On downstream prediction tasks, including in-hospital mortality and 30-day readmission, our sparse representation achieves performance comparable to and in some cases better than baseline approaches, suggesting that the pipeline captures the most clinically salient information.

Furthermore, fine-tuning LLMs (Qwen and MedGemma) on our dataset and evaluating on external datasets demonstrates improved generalization on patient-trajectory–related question answering. For tasks not directly tied to patient trajectories, combining our dataset with task-specific training data yields the strongest results in most cases. Overall, these findings support the effectiveness of our dataset and pipeline across both trajectory-focused and broader clinical question answering settings.

8 Limitations

While our dataset provides a large-scale and structured resource for temporal clinical event modeling, users should be aware of the following limitations:

1. Inconsistencies may exist between the dataset and the original clinical notes.

2. LLM may output hallucinate events that are not actually present in the original note or introduce content from unrelated summaries due to model generalization errors.

3. Timestamp estimation may be imprecise.

We recommend users should apply caution when using this dataset for high-stakes clinical modeling or evaluation tasks, and consider validating subsets of the data with human experts when possible.

References

[1] 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*, pages 72–78.

[2] Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, and et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. In *ACL (System Demonstrations)*.

[3] Steven Bethard and 1 others. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of SemEval 2016*, pages 1052–1062.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners. In *NeurIPS*.

[5] Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 33–40.

[6] Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

[7] Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

[9] William F. Styler IV, Steven Bethard, and 1 others. 2014. Temporal annotation in the clinical domain: The THYME corpus. *Journal of the American Medical Informatics Association*, 21(5):806–815.

[10] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

[11] Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2).

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

[13] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781. Often combined with BM25 in hybrid retrieval.

[14] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR*, pages 39–48. Late-interaction dense retrieval; frequently hybridized with BM25.

[15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

[16] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

[17] Artuur Leeuwenberg and Marie Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246.

[18] Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. 2024. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pages 331–348. Springer.

[19] Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4201–4207.

[20] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[21] Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037.

[22] Long Ouyang, Jeff Wu, Xu Jiang, and et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

[23] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

[24] Rafael Rafailov, Shreyas Sharma, Eric Mitchell, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

[25] Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. 2022. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12):e0000152.

[26] Thomas Searle, Zina Ibrahim, James Teo, and Richard JB Dobson. 2023. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *Journal of Biomedical Informatics*, 141:104358.

[27] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

[28] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

[29] Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

[30] Saran Thaweethai and 1 others. 2023. Development of a Post-Acute Sequelae of SARS-CoV-2 (pasc) cohort in the NIH RECOVER initiative. *medRxiv*. Long COVID cohort; visit-level timelines and symptoms.

[31] Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919.

[32] Leandro von Werra, Younes Belkada, and 1 others. 2022. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

[33] Xuezhi Wang and 1 others. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

[34] Yizhong Wang, Amirreza Mirzaei, and others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *EMNLP*.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

[36] Thomas Wolf and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*.

[37] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

[38] Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

[39] Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61.

[40] Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589.

Table 5: Accuracy (%) of Qwen on MedMCQA by subjects (Part I).

	Anaesthesia	Anatomy	Biochemistry	Dental	ENT	Forensic Med.	Gyn. & Obst.	Medicine	Microbiology	Ophthalmology
Dataset	949	4384	2542	2606	1459	1807	2952	5242	3399	2136
Base	58.27	47.90	57.40	32.46	58.46	62.53	56.37	50.74	54.31	57.91
SubTS	72.08	68.66	80.17	55.53	73.34	76.31	71.31	71.59	76.17	75.23
MCQA	93.57	90.78	94.10	71.26	91.50	95.07	93.60	89.15	93.17	95.08
SubTS+MCQA	93.47	90.67	94.18	71.07	91.57	95.24	93.77	89.18	92.76	95.22

Table 6: Accuracy (%) of Qwen on MedMCQA by subjects (Part II). SPM = Social & Preventive Medicine.

	Orthopaedics	Pathology	Pediatrics	Pharmacology	Physiology	Psychiatry	Radiology	Skin	SPM	Surgery	Unknown
Dataset	920	4549	2433	4214	2643	1320	1327	525	3538	5020	882
Base	56.20	55.62	59.64	51.42	55.35	68.18	62.85	69.90	52.43	54.78	58.05
SubTS	72.83	75.45	74.02	71.76	75.71	80.00	78.52	85.33	71.28	69.38	75.96
MCQA	91.41	92.15	93.55	92.74	92.62	96.67	93.07	97.90	90.87	89.40	92.52
SubTS + MCQA	91.63	92.13	93.59	92.50	92.43	96.59	92.99	97.71	91.10	89.44	92.52

Table 7: Accuracy (%) of MedGemma on MedMCQA by subjects (Part I).

	Anaesthesia	Anatomy	Biochemistry	Dental	ENT	Forensic Med.	Gyn. & Obst.	Medicine	Microbiology	Ophthalmology
Dataset (n)	949	4384	2542	2606	1459	1807	2952	5242	3399	2136
Base	78.50	71.58	81.51	59.25	76.83	81.90	77.34	72.01	80.29	79.92
SubTS	81.14	73.40	83.12	57.21	77.72	85.50	79.34	73.88	81.64	83.90
MCQA	91.25	88.25	91.23	66.23	89.03	94.13	90.48	86.99	90.38	93.54
SubTS + MCQA	91.46	87.80	91.15	66.04	88.62	93.75	90.41	86.82	90.20	93.54

Table 8: Accuracy (%) of MedGemma on MedMCQA by subjects (Part II).

	Orthopaedics	Pathology	Pediatrics	Pharmacology	Physiology	Psychiatry	Radiology	Skin	SPM	Surgery	Unknown
Dataset (n)	920	4549	2433	4214	2643	1320	1327	525	3538	5020	882
Base	77.39	77.64	78.50	77.03	77.90	81.59	80.56	87.81	74.82	75.34	78.00
SubTS	77.83	78.68	81.55	78.57	79.72	83.79	82.37	92.19	77.16	76.45	79.59
MCQA	88.70	90.17	91.45	89.96	89.82	94.39	91.18	97.52	87.62	87.03	90.82
SubTS + MCQA	87.72	90.20	91.66	89.91	89.44	94.39	90.58	97.71	86.94	86.67	89.91

Table 9: Accuracy (%) of Qwen and MedGemma on PubMedQA dataset.

Dataset	Qwen	MedGemma
Base	66.0%	71.33%
SubTS	54.3%	53.66%
PubMedQA	73.7%	72.00%
SubTS + PubMedQA	80.3%	72.00%

Table 10: Accuracy comparison across domains for Qwen and MedGemma under different training setups.

Model	Training data	Clinical	Medical genetics	College biology
Qwen	base	71.7%	74.0%	80.6%
Qwen	subTS	69.43%	74.0%	77.77%
Qwen	MMLU	72.1%	77.0%	79.9%
Qwen	subTS+MMLU	71.69%	73.0%	78.47%
Med-Gemma	base	49.1%	63.0%	55.6%
Med-Gemma	subTS	45.3%	59.0%	52.1%
Med-Gemma	MMLU	54.3%	61.0%	55.6%
Med-Gemma	subTS + MMLU	51.7%	59.0%	55.6%