Large Language Models with Temporal Reasoning for Longitudinal Clinical Summarization and Prediction

Maya Kruse[1], Shiyue Hu[1,2], Nicholas Derby[1,2], Yifu Wu[1], Samantha Stonbraker[1]
Bingsheng Yao[3], Dakuo Wang[3], Elizabeth Goldberg[1], Yanjun Gao[1]
[1]University of Colorado Anschutz Medical Campus
[2]University of Colorado Boulder
[3]Northeastern University

Correspondence: yanjun.gao@cuanschutz.edu

Abstract

Recent advances in large language models (LLMs) have shown potential in clinical text summarization, but their ability to handle long patient trajectories with multi-modal data spread across time remains underexplored. This study systematically evaluates several state-of-the-art open-source LLMs, their Retrieval Augmented Generation (RAG) variants and chain-of-thought (CoT) prompting on long-context clinical summarization and prediction. We examine their ability to synthesize structured and unstructured Electronic Health Records (EHR) data while reasoning over temporal coherence, by re-engineering existing tasks, including discharge summarization and diagnosis prediction from two publicly available EHR datasets. Our results indicate that long context windows improve input integration but do not consistently enhance clinical reasoning, and LLMs are still struggling with temporal progression and rare disease prediction. While RAG shows improvements in hallucination in some cases, it does not fully address these limitations. Our work fills the gap in long clinical text summarization, establishing a foundation for evaluating LLMs with multi-modal data and temporal reasoning. [1]

1 Introduction

Electronic Health Records (EHRs) encapsulate a wide range of multi-modal data, such as vital signs, laboratory results, radiology findings, and free text clinical notes (Mohsen et al., 2022; Li et al., 2022; Belden et al., 2017). Organized across various timestamps, they reflect the dynamic nature of patient care. In clinical settings, particularly for older patients in intensive care units (ICUs) with multiple encounters, chronic conditions, and complex treatment plans, EHRs become especially lengthy and intricate. Clinicians reviewing these long patient histories face significant cognitive burden, often leading to inaccurate decisions such as diagnostic errors (Dymek et al., 2021; Singh et al., 2017). Automated summarization can help improve care continuity and decision-making (Dymek et al., 2021; Adams et al., 2021; Gao et al., 2023a; Laxmisan et al., 2012; Pivovarov and Elhadad, 2015; Liang et al., 2019), but the volume and complexity of EHRs pose challenges, requiring models that handle diverse and evolving clinical data.

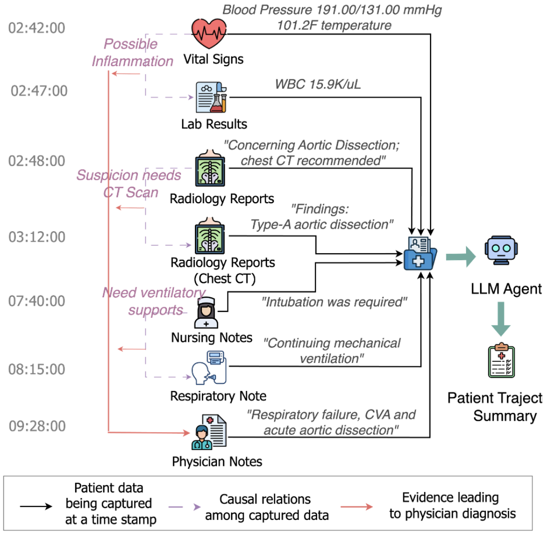
02:42:00
Blood Pressure 191.00/131.00 mmHg
101.2F temperature
Possible Inflammation
Vital Signs
02:47:00
WBC 15.9K/uL
Lab Results
02:48:00
"Concerning Aortic Dissection; chest CT recommended"
Suspicion needs CT Scan
Radiology Reports
03:12:00
"Findings: Type-A aortic dissection"
Radiology Reports (Chest CT)
LLM Agent
07:40:00
Need ventilatory supports
"Intubation was required"
Nursing Notes
08:15:00
"Continuing mechanical ventilation"
Respiratory Note
Patient Trajectory Summary
09:28:00
"Respiratory failure, CVA and acute aortic dissection"
Physician Notes
Patient data being captured at a time stamp
Causal relations among captured data
Evidence leading to physician diagnosis

Figure 1: An illustration of the longitudinal patient trajectory summarization process from multi-modal EHRs. Key causal relationships among medical observations and interventions are highlighted, leading to the final physician diagnosis.

As shown in Figure 1, summarizing longitudinal patient trajectories requires not only *extracting temporally ordered* events from multi-modal data, but also *reasoning* over how earlier findings trigger downstream actions. For instance, abnormal blood pressure and elevated white cell count raise suspicion of inflammation, prompting a chest CT; the CT then confirms an aortic dissection, which leads to intubation and ventilatory support, ultimately consolidated in the physician's diagnostic

[1] Code is available at: https://github.com/LARK-NLP-Lab/longitudinal_clinical_summarization.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20715–20735
November 4-9, 2025 ©2025 Association for Computational Linguistics

note. Capturing these event-to-action links is what we mean by temporal reasoning, and it is critical for preserving clinical causality in the patient trajectory.

Our work evaluates large language models (LLMs) for summarizing complex, longitudinal EHRs, capturing full patient trajectories rather than isolated clinical snapshots. While LLMs have shown promise in clinical NLP tasks (Silcox et al., 2024; Wachter and Brynjolfsson, 2024; Garcia et al., 2024; Adams et al., 2024; Gao et al., 2022), existing evaluations primarily focus on short-context settings and task-specific finetuning. Clinical summarization benchmarks, such as discharge summarization ("Discharge Me"(Xu et al., 2024a)) and diagnosis generation from progress notes ("ProbSum"(Gao et al., 2023a)), evaluate LLMs on isolated segments of patient records rather than full hospital trajectories. While Retrieval-Augmented Generation (RAG) has been applied to clinical tasks like diagnosis prediction and discharge summarization (Xu et al.; Gao et al., 2023b; Lewis et al., 2020; Myers et al., 2024; Lyu et al., 2024), its effectiveness for long, temporally rich clinical narratives remains unclear.

This work addresses two key gaps: (1) current evaluations rarely test long-context LLMs in zero-shot settings, leaving their **inherent capabilities and limitations** in complex medical reasoning unknown; and (2) most clinical benchmarks focus on single time-point summaries, overlooking the demands of **longitudinal patient care**. We systematically evaluate LLMs on full patient trajectories to assess their ability to process temporally evolving, multi-modal clinical data.

In this paper, we focus on two public EHR datasets: Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2020) and EHRShot (Wornow et al., 2023). We evaluate five state-of-the-art LLMs and their RAG variants: Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), Llama3-8B-Instruct (AI@Meta, 2024), Qwen2.5-7B (Yang et al., 2024), DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025) and Llama2-13B-chat-hf (Touvron et al., 2023). Our work advances clinical LLM summarization through the following contributions:

- We reformulate discharge summarization, progress note summarization, and diagnosis classification into new long-context tasks requiring temporal reasoning, aligned with real clinical workflows.
- Our tasks integrate structured and unstructured data across multiple timestamps, enabling analysis of *modality* and *temporal context* effects.
- We compare direct generation, retrieval-augmented generation and chain-of-thought (CoT) prompting (Wei et al., 2022) approaches for handling long clinical documents.

In the absence of benchmarks combining *multimodal inputs*, *temporal structure*, and *clinical workflow alignment*, our study provides a first step toward evaluating LLMs in real-world longitudinal summarization. The results highlight key limitations and suggest directions for more temporally grounded and clinically usable models.

2 Related Work

Clinical text summarization Existing work includes discharge summarization (Xu et al., 2024a; Lyu et al., 2024), diagnosis summarization (Gao et al., 2023a, 2022; Liang et al., 2019), hospital course summarization (Adams et al., 2021, 2024). While these tasks inherently involve multi-document summarization by human physicians, NLP formulations typically treat them as single-document summarization or multi-document summarization with fixed timestamps (e.g. at the time the patient is discharged).

Tabular reasoning LLMs face challenges when reasoning over structured tabular inputs, where tasks often require precise comparisons or aggregations. Zhang et al. (2024) demonstrated the potential of LLMs for table manipulation in real usage scenarios, while Ashury-Tahan et al. (2025) introduced a benchmark showing that models remain sensitive to table formatting. These works highlight open problems in extending LLM reasoning to structured clinical data.

Temporal reasoning Recent studies emphasize the challenges LLMs face in modeling temporal data. Xiong et al. (2024) showed that LLMs can acquire temporal reasoning with a graph framework and targeted prompting, while Tan et al. (2023) provided systematic benchmarks exposing weaknesses and biases in handling temporal entailment and event prediction. Hu et al. (2025) introduced a time-aware agent for temporal knowledge graph question answering, demonstrating the benefit of explicitly encoding temporal constraints. Complementarily, the MenatQA dataset (Wei et al., 2023) highlights systematic failures of LLMs on diverse temporal reasoning tasks. Together, these works underscore the need for dedicated benchmarks and modeling strategies to strengthen LLMs' temporal comprehension, which directly impacts the handling of longitudinal clinical data.

LLMs for long clinical documents Directly processing the long document input can lead to "lost-in-the-middle" problem (Liu et al., 2024). As a result, RAG has been the default when handling long clinical documents, evidenced by its superior performance in diagnosis prediction, discharge summarization and information extraction (Myers et al., 2024; Lyu et al., 2024; Lopez et al., 2025). Due to the task setup, these works have only investigated single modality EHRs with long input length.

Our work addresses this problem by extending RAG and CoT-based approaches to multi-modal EHR summarization and clinical prediction, incorporating both structured (tabular) and unstructured (clinical notes) data to enhance the completeness and accuracy of patient trajectory summaries. We evaluate how well LLMs and their RAG setups can capture and preserve temporal and causal relationships across patient's hospital stay, bridging the gap between isolated document processing and comprehensive longitudinal understanding.

3 Dataset and Tasks

We use two complementary datasets, MIMIC-III and EHRShot, to evaluate LLMs' clinical reasoning. They differ in three key aspects: 1) Modalities: MIMIC-III includes both structured data and clinical notes, while EHRShot contains only structured data. 2) Output type: MIMIC-III tasks focus on generating short summaries of patient progress or discharge status, whereas EHRShot involves classification-based diagnosis prediction. 3) Decision time span: MIMIC-III targets immediate ICU-related summaries, while EHRShot predicts diagnoses within a year of discharge, emphasizing long-term forecasting based on prior visits. Despite these differences, both datasets address longitudinal patient information, requiring models to reason over time and synthesize complex clinical histories.

3.1 MIMIC-III

MIMIC-III is a publicly available dataset comprising de-identified health records from over 40,000 ICU patients. For this study, we focus on a subset of patients with hospital stays exceeding 72 hours to ensure sufficient context for evaluating long-document summarization and temporal reasoning. Unlike prior MIMIC-III studies, our selected cohort emphasizes complex, multi-day ICU stays where accurate summaries are most impactful. This setup allows us to assess LLMs' ability to handle prolonged and evolving clinical narratives.

We use both structured and unstructured data. Specifically, we extract chart events (vital signs, ventilator settings), lab events (e.g., white blood cell counts), input events (e.g., IV infusions, feedings), and medications. These features reflect key aspects of clinical decision-making and support rich, temporally grounded summarization tasks.

Discharge Summarization Discharge summaries are a crucial part of a patient's hospital care process, providing a comprehensive overview of their hospital stay and key clinical events. It is usually written by the physician after the patient is discharged, and contains three main sections: DIAGNOSIS: a list of diagnoses requiring an understanding of the patient's clinical progression up to discharge; BRIEF HOSPITAL COURSE: clinical summary of treatments, interventions, and significant events during hospitalization; DISCHARGE INSTRUCTIONS: post-discharge guidance, including medication plans, dietary recommendations and follow-up care.

Discharge Summarization has been explored in several studies, including (Xu et al., 2024b) and (Ando et al., 2022), often focusing on specific sections like the Brief Hospital Course or discharge instructions. In this paper, we generate *all three sections* by extracting and chronologically ordering structured and unstructured data from a hospital admission. We limit input to the last 24 hours to prevent overloading the LLM while prioritizing the most relevant information for discharge, considering the generally long hospital stays ($\geq$3 days). We also test a 48-hour window to capture a broader clinical context.

We design four input settings to evaluate LLMs on multi-modal and temporal reasoning: NOTE ONLY where input only contains clinical notes, TABULAR ONLY where input only contains tabular data, SHUFFLED TABULAR where we shuffled the tabular data by their timestamps, COMBINED where we combined both notes and tabular data, sorted by their timestamps.

Assessment and Plan (A&P) Generation Daily progress notes are the documents where physicians

Symbol	**Definition**
D_i	Progress note for day i
X_i	Input data used to generate the progress note for day i
Method	**Input Data for Day i**
Baseline	$X_i = \text{EHR}_i$ (No prior progress notes)
Single-Day Context	$X_i = (\text{EHR}_i, D_{i-1})$ (Includes previous day's note)
Multi-Day Context	$X_i = (\text{EHR}_i, D_1, D_2, ..., D_{i-1})$ (Includes all previous notes)

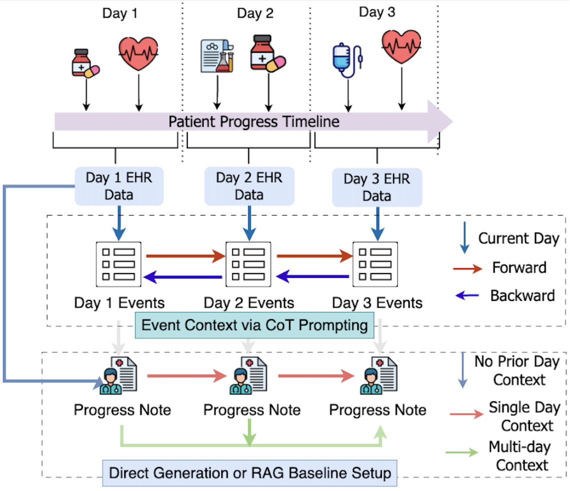

Figure 2: Illustration of the A&P generation workflow (top) and a comparison of input data formulations (bottom).

record diagnoses, treatments, and clinical status, providing key insights into a patient's condition throughout their hospital stay. A progress note typically consists of four sections: Subjective, Objective, Assessment and Plan (Weed et al., 1968; Wright et al., 2014). The Subjective and Objective sections serve as the "evidence" and observation of the patient on that day, with Subjective comprising unstructured free text describing symptoms, status, and treatment, while Objective consists of structured tabular data like lab results and charted values. In contrast, the Assessment and Plan sections capture the physician's reasoning and clinical hypothesis based on this evidence, with diagnoses and treatment plans listed. To assess LLM abilities in summarizing longitudinal data, the Assessment and Plan sections provide a great testbed, requiring integration of information from multiple days, capturing the evolution of the patient's condition, and synthesizing key clinical evidence into a concise yet comprehensive summary.

The setup of the task is illustrated in Figure 2, which explains the three input methods: the NO PRIOR CONTEXT (BASELINE) method uses only the current day's structured EHR data, the SINGLE-DAY CONTEXT method adds the previous day's progress note for limited historical context, and the MULTI-DAY CONTEXT method incorporates all prior progress notes for a more comprehensive patient history. For day i, the input includes progress notes up to $i - 1$ (depending on the method), but the note for i is excluded and compared to the generated version. This setup reflects the real clinical workflow, where physicians reference prior notes when writing new ones.

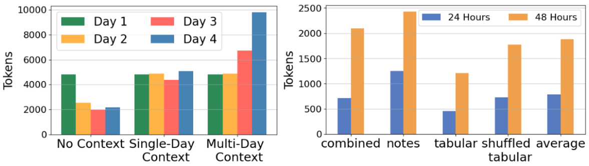

Figure 3: The token length of input data used for A&P generation across days and methods (left), as well as that of input data used for Discharge Summarization across time windows and modalities(right).

How long is the input in MIMIC? Figure 3 shows input lengths across tasks and settings. For discharge summarization, inputs average 500 tokens (24-hour window) and 1,940 tokens (48-hour window). In contrast, the Assessment and Plan (A&P) generation task involves much longer inputs due to the inclusion of prior progress notes and structured data. The No Context baseline starts with a relatively high token count (3,125 tokens on Day 1) but remains the shortest overall. Single-Day Context, which adds the previous day's note, increases input length to 5,375 tokens. Multi-Day Context, which accumulates all prior notes, shows the steepest token growth, reaching nearly 10,000 tokens by Day 4 and averaging 6,875 tokens.

3.2 Diagnosis Prediction from EHRShot

EHRShot (Wornow et al., 2023) is a longitudinal dataset comprising fully structured EHR data from 6,739 patients at Stanford Medicine, with over 41 million clinical events. In contrast to MIMIC-III, which contains a mix of structured and unstructured data, EHRShot offers a clean, structured-only setting, allowing for controlled evaluation of temporal reasoning in longitudinal patient records.

We include EHRShot to complement the MIMIC-based summarization tasks in two important ways. First, EHRShot focuses on future prediction—each task requires predicting whether a patient will receive one of six diagnoses (e.g., HYPERTENSION, PANCREATIC CANCER, ACUTE MI) within one year after discharge. This differs from our MIMIC tasks, which emphasize retrospective summarization of observed hospital stays. Second, its purely structured nature allows us to isolate models' abilities to handle temporal patterns in tabular data without the added complexity of clinical language. Together, these contrasts offer a broader assessment of LLM capabilities in modeling temporal EHR information across different modalities and predictive targets.

The six diagnosis tasks vary in prevalence from common conditions like HYPERLIPIDEMIA (31.85%) to rarer ones like CELIAC (3.46%). Inputs have a mean token length of 1,989 (σ: 1,270), with a median of 1,851, and an average of 74 unique measurements per patient. More statistics are provided in Appendix A.2.

First entry: Urea Nitrogen is 26 mg/dL. Hematocrit is 34.20%. Hemoglobin is 12.20 g/dL. 12 minutes later: Glucose is 214 mg/dL. Lactate is 2.40 mmol/L. **3 minutes later:** Hemoglobin is 13.00 g/dL. Lactate is 2.40 mmol/L. pO2 is 441 mm Hg. **1 hour later**: 100.00 ml of 0.9% Normal Saline is administered. Neosynephrine-k is administered. **49 minutes later:** Radiology note: 12:48 PM CHEST (PORTABLE AP) Clip Reason: line placement, r/o PTx Admitting Diagnosis: HEAD BLEED; MEDICAL CONDITION: 68-year-old man s/p MVA significant head trauma, intubated s/p r subclavian triple lumen placement...

Table 1: An example of a patient's compiled data, including relative timestamp (minutes and hours since previous recorded data), structured data converted into narrative format as well as unstructured note data.

3.3 Representing multi-modal, longitudinal patient data

We convert structured EHR data into natural language to enable LLM-based summarization, following prior table-to-text approaches (Gao et al., 2024; Yu et al., 2023). As in Table 1, each measurement is verbalized using simple templates (e.g., [Measurement] is [Value][Unit]), grouped by timestamp, and temporally ordered using relative time references. Medications and input events follow a similar format (e.g., "is administered").

To reduce redundancy from repeated or copy-pasted entries, we deduplicate records across modalities and retain only the most recent entry when multiple identical values appear within a short time window. Clinical notes are filtered similarly. For EHRShot, structured inputs are divided into six hour chunks and repetitive phrases are removed. Appendix A.1.

4 Methods

We evaluate five LLMs with varying context capacities. LLaMA2-13B and Mistral-7B support up to 4K tokens, while LLaMA3-8B allows for 8K. To test long-context capabilities, we include Qwen2.5-7B and DeepSeek-R1 (32B), both of which support input lengths up to 128K tokens. While we did explore biomedical LLMs (PMC-LLaMa, BioGPT, BioMistral), their performance was extremely poor, so we did not include them in this paper.

4.1 Approaches

We compare a baseline direct generation, structured RAG and CoT event extraction approach. The key difference between these approaches lies in how they handle long temporal contexts. Direct generation processes the entire input at once but may suffer from the lost-in-the-middle problem, where relevant information buried in long documents is overlooked. In contrast, RAG segments temporal information into retrievable chunks, which helps manage long sequences but may disrupt temporal dependencies between events. For CoT event extraction, the long context is first compressed into a series of clinical events, over which the model then summarizes.

1). Direct Generation. The patient's chronological data, formatted according to the specific task requirements, is provided to the model as a direct input without any additional structural modifications. This serves as a baseline to assess the capability of current state-of-the-art language models in processing and generating clinical summaries from raw sequential data, without the aid of task-specific adaptations or architectural enhancements.

2). RAG. For this setup, we choose the same selection as models used for direct generation, with the exception of DeepSeek, as it is considerably larger than the other LLMs (32B parameters vs 7-13B) yet fails to outperform these smaller models. This combination of high computational cost and suboptimal performance led us to exclude DeepSeek for RAG. Myers et al. (2024) studied the quality of embeddings in medical RAG and found BGE (Xiao et al., 2023) yielded the highest performance. Thus, we adopt BGE embeddings and apply them to all LLM RAG setups in our work. Additionally, the queries presented in the paper are slightly modified and used to perform query optimization. Hyperparameter optimization is carried out on the standard RAG hyperparameters chunk size, chunk overlap and top-k retrieved documents.

3). Event Extraction via CoT prompting. As an alternative to retrieval or direct generation, we introduce a CoT prompting approach that incorporates a clinical reasoning step before summarization. LLMs are prompted to first identify temporally ordered clinical events from the input data, which are appended to the original input as structured context for generation. This intermediate step emphasizes salient clinical signals and reduces input noise, helping the model focus on meaningful patterns.

We develop our prompt with guidance from recent studies showing that LLMs perform better on clinical tasks when given clear, structured instructions tailored to the specific goal. For example, Wang et al. (2023) finds that prompts targeting specific event types like symptoms, lab results, treatments, and medical decisions help models focus on clinically meaningful information and improve event detection. Yuan et al. (2023) also shows that when complex tasks are broken into smaller reasoning steps, especially for understanding how events are ordered over time, models tend to produce more consistent and accurate outputs. Figure 4 illustrates the CoT prompt design for A&P generation task.

We experiment with seven temporal input configurations for event extraction: using only the current day's data, combining it with previous days (Forward), or with subsequent days (Backward), each spanning 1, 3, or all available days. We apply these settings to a development batch of 20 patients and found that, within each respective group of experiments, the Forward setting using all previous days and the Backward setting using one following day produced the best results. Based on these findings, we report performance under three representative temporal contexts: +0 (Current Day Only, baseline), +N (Forward Context with all prior days), and −1 (Backward Context with one following day).

Chain-of-Thought Prompt for ICU Daily Event Extraction

ICU DAILY EVENT EXTRACTION TASK
Analyze this structured ICU data by identifying critical clinical events, paying special attention to numerical values and their progression. Only report values showing meaningful change or clinical significance. For repeated values, mention only those demonstrating changes.
{Input Text}
Identify (with direct references to data points when possible):

1. Major symptoms or changes (new, worsening, improving) – Specify relevant numerical changes.
2. Critical test results (labs, imaging, etc.) – Highlight significant abnormal or normal values.
3. Important treatments or interventions – Clearly link to the preceding clinical data.
4. Significant care team decisions – Support with relevant clinical data.
5. Major medical decisions or diagnoses – Reference pertinent clinical observations.

Response Format:
### Day X Key Events ###
- [Time] | [Event Description] (Explanation for identifying this event)

Example Model Output

### Day 3 Key Events ###
- **2178-02-11 09:50:00** | Blood pressure readings of 88/46 mmHg (This low blood pressure reading is concerning and may require additional fluid resuscitation or adjustment of vasoactive medications.)

Figure 4: Chain-of-Thought prompting template and example output used for ICU daily event extraction from structured EHR data for A&P generation.

Experiment settings. For all LLMs, we run their 8-bit quantized version. We set the output token length as 1,000, but almost all tasks output is significantly shorter than this limit. For RAG setup, we used Langchain (Chase, 2022) with Faiss (Douze et al., 2024) for semantic retrieval in a vector database. We perform hyperparameter tuning with chunk size between [250, 750], top k between [10, 50], chunk overlap between [50,200]. The detailed results of hyperparameter searching are in Appendix A.4. All experiments are run on 2 H100 94GB GPUs.

4.2 Evaluation

On MIMIC, we report standard summarization evaluation metrics that capture string overlap and semantics similarity. ROUGE-L (Lin, 2004) measures string overlap, while BERTScore (Zhang et al., 2019) assesses maximum token pairwise similarity. We use SapBERT as the backend for BERTScore due to its superior performance in biomedical entity representation (Liu et al., 2020). On EHRShot, given that the diagnosis prediction tasks are essentially binary classification, we report macro-averaged accuracy and F-scores.

5 Results

The results are organized by task: Discharge Summarization and Assessment and Plan Generation

Setting	Model	ROUGE-L	BERTScore
Direct Gen	Mistral	16.28 ±12.83	65.23 ±13.38
	Llama3	11.82 ±14.42	55.35 ±22.48
	Qwen	15.28 ±5.38	63.98 ±13.50
	DeepSeek	13.51 ±5.04	63.64 ±13.86
	Llama2	15.34 ±5.90	63.82 ±11.44
RAG	Mistral	15.04 ±2.34	65.89 ±5.69
	Llama3	15.90 ±2.00	64.38 ±4.81
	Qwen	**17.91** ±2.35	65.25 ±4.77
	Llama2	16.98 ±1.59	**67.20** ±6.21

Table 2: Results on discharge summarization, comparing the direct generation and RAG approaches across all models.

Section	Method	ROUGE-L	BERTScore
Dx	Direct Gen	**3.42** ± 4.34	**50.07** ± 10.78
	CoT Prompt	2.95 ± 3.48	48.46 ± 11.00
HC	Direct Gen	**12.28** ± 3.13	**62.51** ± 5.95
	CoT Prompt	9.98 ± 3.84	59.60 ± 7.32
DI	Direct Gen	**12.07** ± 3.90	60.05 ± 9.09
	CoT Prompt	10.83 ± 4.58	**60.52** ± 7.86

Table 3: Mistral performance on the three sections of discharge summarization (Dx: Diagnosis, HC: Hospital course, DI: Discharge Instruction), comparing direct generation approach with the event extraction CoT.

performed on MIMIC-III and Diagnosis Prediction on EHRShot. Additionally, prompt optimization was carried out on all tasks, using a small sample set (Appendix A.6).

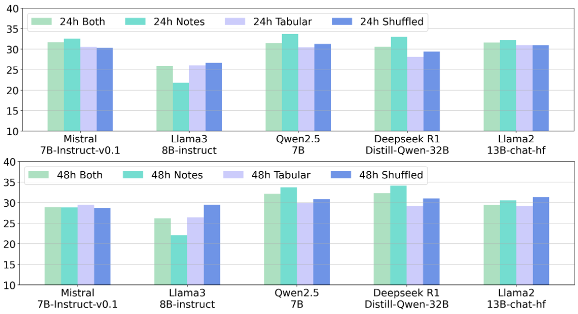

Figure 5: Average f1 results across modalities and time windows (on direct generation)

Discharge Summarization. Table 2 presents results from comparison between the Direct Generation and RAG approaches. Qwen achieves the highest ROUGE-L score of 17.91 using RAG. In general, most LLMs record a performance increase in their RAG variants, the biggest of which is Llama3's, with an increase of 4.08 on ROUGE-L. Mistral is the exception, with a minor performance decrease of -1.24. DeepSeek reports mediocre performance despite its larger size.

Table 3 shows the section-wise analysis of Mistral comparing direct generation and CoT methods, where CoT event extraction performs slightly worse than direct generation in most sections. Other LLMs exhibit similar trends, so we report Mistral only for brevity. In Appendix 8, we provide more section-wise analysis on Discharge Summ task. Results further confirm the challenge of LLMs reasoning over patient trajectories.

Figure 5 shows the effects of the four modalities (notes + tabular data, notes only, tabular only and shuffled tabular) and two context window sizes (24 and 48 hours) on performance. For the 24 hour window, the pure note modality dominates, performing best on Mistral, Qwen, DeepSeek and Llama2. Llama3 is the outlier and reports the best performance on the combined and shuffled tabular modalities. In the 48h window, trends are less clear, but most models still favor notes.

Temporal order appears unhelpful, as shuffled tabular data performs slightly better than chronological tabular data, likely because the data is near the discharge state, where patients' clinical conditions stabilize, leading to fewer changes over time.

A& P Generation. Table 4 shows that adding prior context (Single- or Multi-Day) improves performance across all models, particularly for ROUGE scores. Llama3 performs best on direct generation, especially in Single- and Multi-Day Contexts, but is outperformed by Qwen and Mistral on RAG. However, Multi-Day Context does not always yield performance gains over Single-Day, suggesting models struggle with longer patient histories. RAG generally improves performance, particularly for Mistral and Qwen, but its impact varies across models: Llama3 performs better without retrieval.

This task specifically requires clinical reasoning over past captured data (evidence) to summarize patient progression and plan for treatment, yet the relatively low ROUGE and BERTScore values indicate that models still struggle with temporal reasoning and integrating historical context effectively.

Table 5 presents results of the event extraction CoT method. CoT did not yield improvements over direct generation or RAG approaches across ROUGE-L or BERTScore. This suggests that, while the event extraction step offers interpretability and control, it may not directly enhance generation quality when applied in a pipeline without further model adaptation. We view this as a useful diagnostic approach that could be further refined or integrated with instruction tuning in future work.

Diagnosis Prediction. We evaluate Mistral and

Setting	*Direct Generation*						*RAG*					
	Mistral		Llama3		Qwen		Mistral		Llama3		Qwen	
	RL	BS	RL	BS	RL	BS	RL	BS	RL	BS	RL	BS
No Prior	**17.70**±3.79	68.16±6.91	16.60±3.39	72.01±6.04	15.15±3.53	72.17±4.51	17.15±5.60	68.57±6.16	15.48±3.60	**73.78**±4.50	15.60±3.60	72.16±4.63
Single-Day	26.11±11.68	71.63±10.86	**33.16**±10.68	80.10±8.93	20.74±4.19	78.01±4.95	29.50±12.46	74.23±7.79	22.09±5.51	78.80±4.96	25.42±8.08	**79.27**±5.37
Multi-Day	23.32±13.05	72.41±9.47	**32.42**±12.14	**80.42**±8.81	21.60±5.23	78.41±4.81	27.40±10.08	73.55±8.46	21.42±6.81	78.96±5.42	25.36±7.50	80.28±4.68

Table 4: Performance of Direct Generation and RAG methods on A&P generation across models and input settings. RL = ROUGE-L score; BS = BERTScore computed using SapBERT embeddings.

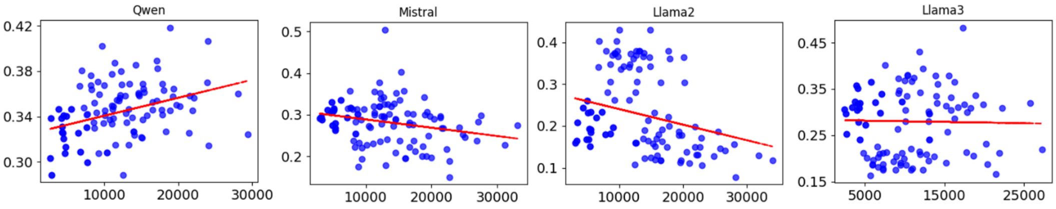

Figure 6: Correlation of token length and average performance across all tasks and metrics

Model	Metric	Current Day	Forward+N	Backward-1
Mistral	ROUGE-L BERTScore	19.85±10.87 68.63±11.56	20.82±12.02 69.69±10.82	20.10±11.15 69.35±9.67
Llama3	ROUGE-L BERTScore	**31.56**±16.56 **77.89**±12.48	31.34±16.93 76.17±14.74	**28.40**±14.78 74.66±16.31
Qwen	ROUGE-L BERTScore	20.41±4.39 77.44±4.17	20.69±4.65 77.61±4.27	20.70±4.70 **77.10**±4.40

Table 5: Event extraction CoT results on A&P generation task across settings.

		Direct Gen		RAG	
		Mistral	Qwen	Mistral	Qwen
Acute MI	Accuracy	52.84	62.72	63.03	65.50
	F1	33.45	2.58	11.83	16.87
Celiac Disease	Accuracy	62.96	96.05	95.04	95.30
	F1	2.60	0.00	0.00	9.52
Hyperlipidemia	Accuracy	58.02	68.15	66.09	63.03
	F1	30.33	1.53	12.74	7.45
Hypertension	Accuracy	64.44	69.38	68.81	68.81
	F1	26.53	6.06	8.7	8.7
Lupus	Accuracy	66.42	95.06	94.79	94.81
	F1	10.53	0.00	0.00	0.00
Pancreatic Cancer	Accuracy	73.83	78.52	78.86	79.05
	F1	13.11	2.25	0.00	0.00

Table 6: Results on the binary diagnosis prediction task using EHRShot data (macro-average).

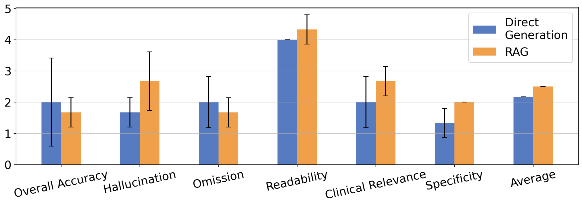

Figure 7: Expert review scores for Qwen direct generation and RAG on Discharge Summarization

Qwen for this task based on their prior performance (results shown in Table 6). Both models prioritize majority class predictions, leading to inflated accuracy but severely low F1-scores. RAG improves accuracy but does not meaningfully enhance recall or F1-score.

The task remains highly challenging, especially for rare diseases, where models struggle to predict positive cases. Qwen achieves high accuracy but an F-score of zero for celiac disease and lupus, reflecting severe class imbalance. It mostly predicts negatives, occasionally misclassifying positives but failing to identify true cases. Given that 96.54% of celiac and 95.56% of lupus cases are true negatives, its accuracy (96.05% and 95.06%) is barely lower than predicting all negatives.

Overall Performance Figure 6 illustrate how input length impacts model performance across different LLMs, considering their varying context windows. Qwen (128K context), shows a slight positive correlation, suggesting that longer inputs may improve its performance. In contrast, Mistral and Llama2 (both 4K context), decline as input length increases, likely due to exceeding their optimal processing capacity. Llama3 (8K context), remains stable with minimum impact from the length. Models with shorter contexts struggle with longer inputs, while larger-context models handle them better, though benefits remain inconsistent.

6 Discussion

Another critical aspect to consider is running time and memory usage. All 7B LLMs require approximately 15GB of GPU RAM for direct generation. RAG demands significantly more memory, peaking at 32–33GB, nearly twice that of direct generation. For Qwen, it takes 18-20 minutes to run RAG on one progress note with prior context input, while direct generation takes 10 minutes on average. On EHRShot where input is shorter, diagnosis predic-

Approach	Task	CIT	ACC	THR	USE	ORG	CMP	SUC	SYN
Direct Gen	DS	1.00±0	1.82±1.21	1.42±0.64	1.70±0.93	2.48±0.99	4.08±0.75	2.96±0.92	3.5±0.71
	A&P	2.13±0.89	3.19±1.15	2.66±0.87	3.51±0.88	3.46±0.86	3.84±0.8	3.03±0.92	3.76±0.66
RAG	DS	1.06±0.24	1.68±0.88	1.52±0.79	1.66±0.94	2.4±1.01	3.92±0.78	2.78±0.76	3.5±0.71
	A&P	2.35±1.02	3.96±0.98	3.46±0.76	4.07±0.81	3.71±0.9	4.11±0.74	3.00±0.91	3.61±0.7

Table 7: esults of the PDSQI-9 evaluation on discharge summarization (DS) and Assessment and Plan generation (A&P). Abbreviations: CIT = citations, ACC = accuracy_extractive, THR = thoroughness, USE = usefulness, ORG = organization, CMP = comprehensibility, SUC = succinctness, SYN = synthesis_abstraction. Definitions of scoring criteria could be found in (Croxford et al., 2025).

tion takes just 3–4 seconds. For other LLMs with smaller context window, they run quicker.

Having established the computational feasibility of these models, we next turn to how their outputs are evaluated for clinical quality.

LLM-As-Judge The Provider Documentation Summarization Quality Instrument (PDSQI-9) is a clinically validated framework that specifies nine dimensions of summarization quality for the evaluation of LLM summarization (Croxford et al., 2025). In addition to clinician evaluation, Croxford et al. (2025) implemented and validated an LLM-as-Judge based on this framework, using GPT-o3, and showed that the approach is strongly correlated with clinician ratings. In our study, we adapt their released rubric and prompting protocol within our HIPAA-compliant Azure OpenAI environment. We did not introduce any task-specific tuning beyond template filling. This setup was used to score Qwen-generated discharge summaries and A&P sections and served as a scalable complement to traditional automated metrics. Table 7 shows the evaluation results: RAG generally outperforms Direct Generation, particularly in accuracy, thoroughness, and usefulness, with the strongest results seen in the A&P task. Across both approaches, A&P summaries score higher than DS, suggesting that model performance improves when the task involves structured, focused inputs (A&P) rather than broad, comprehensive records (discharge summaries). Comprehensibility is consistently strong, indicating good clarity, while succinctness remains middling, showing redundancy issues. The weakest category across all conditions is citations, where scores remain low, highlighting a major gap in source attribution. Overall, RAG with A&P produces the most balanced and reliable summaries.

Expert error analysis To provide a more nuanced perspective, we also consulted with a senior board-certified Emergency Department physician for qualitative review. This assessment was *not* intended to yield quantitative results, but rather to capture how physicians perceive the generated summaries and to identify areas for improvement beyond automated scores. We focus specifically on discharge summarization and sample 10 pairs of RAG and Direct Generation output from Qwen, as it reached relatively high performance on automated metrics. The evaluation criteria were adapted from PDSQI-9, but streamlined and combined with categories from prior clinical text evaluation studies: OVERALL ACCURACY (factual correctness), HALLUCINATION,OMISSION, READABILITY, CLINICAL RELEVANCE and SPECIFICITY (Singhal et al., 2023; Xu et al., 2024a; Ben Abacha et al., 2023; Aljamaan et al., 2024; Croxford et al., 2025; Williams et al., 2024). The results of this analysis are given in Figure 7, and the annotation guidelines can be found in Table 20. While PDSQI-9 consolidates certain aspects (e.g., "accuracy" combines factual correctness and hallucination), our interest lies in examining these dimensions at a more granular level. For example, we evaluate OVERALL ACCURACY and HALLUCINATION separately to better capture different error types.

Overall, RAG performs slightly better, with fewer hallucinations and more clinically relevant summaries. However, both methods still struggle—often prioritizing less relevant diagnoses, oversimplifying summaries, and failing to discard outdated or disproven diagnoses. These issues highlight ongoing challenges with temporal reasoning.

7 Conclusion

We evaluated LLMs on long-context clinical summarization and prediction using two public EHR datasets. Current models struggle with accurate summarization and temporal reasoning, making it hard to interpret medical event sequences. While RAG offers some improvement, overall performance remains inadequate, highlighting the need for further progress.

Acknowledgments

This work is supported by U.S. National Library of Medicine R00 LM014308.

Limitations

This study focused on current LLMs' capabilities on the task of clinical summarization for long context documents including temporal information. We evaluated several different models, including Qwen, Llama 3 and DeepSeek, but we acknowledge that this selection is by no means comprehensive and could have limited our analysis. It also does not contain any closed-source LLMs, as the use of these models with MIMIC and EHRShot data is prohibited by Data Use Agreements. Due to the dearth of publicly available EHR datasets, we were only able to include data from two sources: MIMIC-III and EHRShot.

Our human evaluation was constrained by real-world clinical scheduling limitations, allowing us to consult only one emergency department (ED) physician. Furthermore, we did not formally validate our proposed survey instrument, though the survey questions were aggregated from established prior work (see Table 20 for the literature these questions came from). Our goal is to leverage domain experts to better understand LLM limitations, and we plan to expand human evaluation efforts in future studies to provide a more comprehensive assessment.

Ethical Statement

This study utilizes de-identified patient data from publicly available datasets (MIMIC-III and EHRShot), ensuring that no identifiable patient information is used. As a result, our work does not involve human subjects research and poses no risk to patient privacy or confidentiality.

Additionally, our study is purely retrospective and computational, focusing on evaluating LLMs for clinical summarization. The models analyzed do not interact with real-world clinical workflows and are not used for actual medical decision-making. Therefore, there is no potential harm to patients or healthcare providers as a result of this research.

Our goal is to assess and improve LLM capabilities for handling complex clinical data, with the long-term aim of developing safe, explainable, and effective AI tools to support healthcare professionals in the future. However, we acknowledge that applying existing LLMs in clinical workflows carries real risks, including privacy leakage, algorithmic biases, and the potential for incorrect decisions. Addressing these challenges is critical for ensuring the safe and ethical deployment of AI in healthcare.

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4794. NIH Public Access.

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2024. Speer: Sentence-level planning of long clinical summaries via embedded entity retrieval. *arXiv preprint arXiv:2401.02369*.

AI@Meta. 2024. Llama 3 model card.

Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, et al. 2024. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics*, 12(1):e54345.

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Is artificial intelligence capable of generating hospital discharge summaries from inpatient records? *PLOS Digital Health*, 1(12):e0000158.

Shir Ashury-Tahan, Yifan Mai, Ariel Gera, Yotam Perlitz, Asaf Yehudai, Elron Bandel, Leshem Choshen, Eyal Shnarch, Percy Liang, Michal Shmueli-Scheuer, et al. 2025. The mighty torr: A benchmark for table reasoning and robustness. *arXiv preprint arXiv:2502.19412*.

Jeffery L Belden, Richelle J Koopman, Sonal J Patil, Nathan J Lowrance, Gregory F Petroski, and Jamie B Smith. 2017. Dynamic electronic health record note prototype: seeing more by showing less. *The Journal of the American Board of Family Medicine*, 30(6):691–700.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.

Harrison Chase. 2022. Langchain. Accessed: 2025-02-14.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Miranda Schnier, Kyle Burton, Cris Ebby, Jillian Gorski, et al.

2025. Development and validation of the provider documentation summarization quality instrument for large language models. *Journal of the American Medical Informatics Association*, 32(6):1050–1060.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Christine Dymek, Bryan Kim, Genevieve B Melton, Thomas H Payne, Hardeep Singh, and Chun-Ju Hsiao. 2021. Building the evidence-base to reduce electronic health record–related clinician burden. *Journal of the American Medical Informatics Association*, 28(5):1057–1061.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. 2023a. Overview of the problem list summarization (probsum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 461. NIH Public Access.

Yanjun Gao, Timothy Miller, Dongfang Xu, Dmitriy Dligach, Matthew M Churpek, and Majid Afshar. 2022. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of COLING. International Conference on Computational Linguistics*, volume 2022, page 2979. NIH Public Access.

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? *arXiv preprint arXiv:2408.11854*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Patricia Garcia, Stephen P Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, et al. 2024. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Network Open*, 7(3):e243201–e243201.

Qianyi Hu, Xinhui Tu, Cong Guo, and Shunping Zhang. 2025. Time-aware ReAct agent for temporal knowledge graph question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6013–6024, Albuquerque, New Mexico. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A

Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, pages 1–10.

Archana Laxmisan, Allison B McCoy, Adam Wright, and Dean F Sittig. 2012. Clinical summarization capabilities of commercially-available and internally-developed electronic health records. *Applied clinical informatics*, 3(01):80–93.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Rui Li, Fenglong Ma, and Jing Gao. 2022. Integrating multimodal electronic health records for diagnosis prediction. In *AMIA Annual Symposium Proceedings*, volume 2021, page 726.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, et al. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Mengxian Lyu, Cheng Peng, Daniel Paredes, Ziyi Chen, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. Uf-hobi at "discharge me!": A hybrid solution for discharge summary generation through prompt-based tuning of gatortrongpt models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 685–695.

Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. 2022. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1):17981.

Skatje Myers, Timothy A Miller, Yanjun Gao, Matthew M Churpek, Anoop Mayampurath, Dmitriy Dligach, and Majid Afshar. 2024. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. *Journal of the American Medical Informatics Association*, page ocae308.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Christina Silcox, Eyal Zimlichmann, Katie Huber, Neil Rowen, Robert Saunders, Mark McClellan, Charles N Kahn III, Claudia A Salzberg, and David W Bates. 2024. The potential for artificial intelligence to transform healthcare: perspectives from international health leaders. *NPJ Digital Medicine*, 7(1):88.

Hardeep Singh, Gordon D Schiff, Mark L Graber, Igho Onakpoya, and Matthew J Thompson. 2017. The global burden of diagnostic errors in primary care. *BMJ quality & safety*, 26(6):484–494.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Robert M Wachter and Erik Brynjolfsson. 2024. Will generative artificial intelligence deliver on its promise in health care? *Jama*, 331(1):65–69.

Sijia Wang, Mo Yu, and Lifu Huang. 2023. The art of prompting: Event detection based on type specific prompts. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Lawrence L Weed et al. 1968. Medical records that guide and teach. *N Engl J Med*, 278(11):593–600.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. *arXiv preprint arXiv:2310.05157*.

Christopher YK Williams, Brenda Y Miao, Aaron E Kornblith, and Atul J Butte. 2024. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature Communications*, 15(1):8236.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. 2023. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models.

Adam Wright, Dean F Sittig, Julie McGowan, Joan S Ash, and Lawrence L Weed. 2014. Bringing science to medicine: an interview with larry weed, inventor of the problem-oriented medical record. *Journal of the American Medical Informatics Association*, 21(6):964–968.

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

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024a. Overview of the first shared task on clinical text generation: RRG24 and “discharge me!”. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, et al. 2024b. Overview of the first shared task on clinical text generation: Rrg24 and “discharge me!”. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Unified language representation for question answering over text, tables, and images. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4756–4765.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiaokang Zhang, Sijia Luo, Bohan Zhang, Zeyao Ma, Jing Zhang, Yang Li, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, et al. 2024. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*.

A Appendix

A.1 More Preprocessing Details

In addition to the preprocessing steps outlined in section 3.3, we attempt, as far as possible, to only include laboratory measurements and vital sign values that fall outside the normal range. These values are more likely to indicate a problem the patient is facing and are thus more salient. For lab measurements, we only keep those values flagged as "abnormal". For chart or vital sign values, there exists a warning flag, which is either 0 or 1. However, some values have an NaN value instead. We exclude values with the flag set to 0, but keep those with 1 or NaN, since it is unknown whether the value corresponding to the NaN type is abnormal or not without a deeper medical analysis.

A.2 EHRShot Statistics

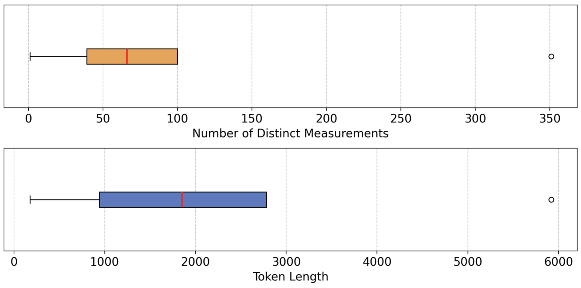

Figure 8: Additional statistics on the EHRShot dataset. Number of distinct measurements given above, with the token length distribution below.

Figure 8 reports detailed statistics of EHRShot cohort, regarding the input token length and number of distinct clinical measurements. The prompt token lengths exhibit a wide range, with a mean of 1,989 tokens and a standard deviation of 1,270 tokens. The distribution is right-skewed, with a minimum of 178 tokens and a maximum of 5,919 tokens, while the median (50th percentile) is 1,851 tokens.

Model	Sec.	CUI	ROUGE-L	BERTScore
Llama2	Dx	3.87 ± 5.23	4.01 ± 3.17	52.21 ± 11.59
	HC	7.07 ± 4.89	12.26 ± 3.50	61.99 ± 6.12
	DI	**10.28** ± 7.67	**13.02** ± 5.18	**63.04** ± 11.68
Mistral	Dx	5.46 ± 10.07	3.42 ± 4.34	50.07 ± 10.78
	HC	8.64 ± 5.98	**12.28** ± 3.13	**62.51** ± 5.95
	DI	**9.09** ± 6.90	12.07 ± 3.90	60.05 ± 9.09

Table 8: Discharge Summarization: Llama2 and Mistral Performance Across Sections: Diagnosis (Dx), Hospital Course (HC), and Discharge Instructions (DI).

For distinct measurement types per patient, the data is also highly variable, with a mean of 74 and a standard deviation of 48. The number of distinct measurements ranges from 1 to 351, with a median of 66 and an interquartile range from 39 (25th percentile) to 100 (75th percentile).

Both distributions highlight significant variation in input complexity, reinforcing the need for models to handle long-context dependencies and multimodal data effectively.

A.3 More results on direct generation, RAG and CoT performance comparison

Table 7 compares Llama2 and Mistral models across three discharge sections: Diagnosis (Dx), Hospital Course (IIC), and Discharge Instructions (DI). Table 8 reports performance on A&P generation, comparing direct generation and RAG approaches for Mistral, Llama3, Qwen, and DeepSeek. Table 9 focuses on Mistral's discharge summarization, comparing direct generation with a Chain-of-Thought prompting method.

A.4 Hyperparameter searching results

In this section, we report our hyperparameter tuning on the RAG setup for all LLMs. Table 11 covers the hyperparameter selection. All tuning is done on Discharge Summarization and A&P generation tasks using a small held-out set (n=5). The best set of parameters is used for official experiments reported in the main text. The results of hyperparameter finetuning on Discharge Summarization and A&P generation are given in Tables 12 and 13 respectively.

A.5 Event extraction COT prompt for discharge summarization

The prompt used for event extraction on the discharge summarization task is provided in Figure 9.

Model	Metric	No Prior	Single-Day	Multi-Day
		Direct Generation		
Mistral	CUI	21.60 ± 6.47	34.06 ± 12.15	30.55 ± 13.48
	ROUGE	17.70 ± 3.79	26.11 ± 11.68	23.32 ± 13.05
	BERTScore	68.16 ± 6.91	71.63 ± 10.86	72.41 ± 9.47
	Average	35.82 ± 5.72	43.93 ± 11.56	42.09 ± 12
Llama3	CUI	21.83 ± 6.51	45.35 ± 10.45	43.55 ± 12.00
	ROUGE	16.60 ± 3.39	33.16 ± 10.68	32.42 ± 12.14
	BERTScore	72.01 ± 6.04	80.10 ± 8.93	80.42 ± 8.81
	Average	36.81 ± 4.80	**52.87** ± 10.02	**52.13** ± 10.98
Qwen	CUI	21.84 ± 5.68	33.8 ± 6.63	33.96 ± 6.56
	ROUGE	15.15 ± 3.53	20.74 ± 4.19	21.6 ± 5.23
	BERTScore	72.17 ± 4.51	78.01 ± 4.95	78.41 ± 4.81
	Average	36.38 ± 4.57	44.18 ± 5.26	44.65 ± 5.53
DeepSeek	CUI	22.58 ± 6.63	34.45 ± 9.26	34.15 ± 10.00
	ROUGE	16.16 ± 3.15	21.84 ± 5.65	21.07 ± 5.42
	BERTScore	74.97 ± 3.66	78.29 ± 4.61	78.35 ± 5.12
	Average	**37.90** ± 4.48	44.86 ± 6.51	44.52 ± 9.51
		RAG		
Mistral	CUI	22.65 ± 7.94	35.01 ± 12.40	34.17 ± 10.57
	ROUGE	17.15 ± 5.60	29.50 ± 12.46	27.40 ± 10.08
	BERTScore	68.57 ± 6.16	74.23 ± 7.79	73.55 ± 8.46
	Average	36.12 ± 6.57	46.24 ± 10.88	45.04 ± 9.70
Llama3	CUI	20.33 ± 5.59	32.02 ± 5.93	31.60 ± 8.38
	ROUGE	15.48± 3.60	22.09 ± 5.51	21.42 ± 6.81
	BERTScore	73.78 ± 4.50	78.80 ± 4.96	78.96 ± 5.42
	Average	36.53 ± 4.56	44.3 ± 5.47	43.99 ± 6.87
Qwen	CUI	22.07± 6.23	37.3 ± 8.83	37.21± 8.15
	ROUGE	15.6 ± 3.6	25.42 ± 8.08	25.36 ± 7.5
	BERTScore	72.16 ± 4.63	79.27 ± 5.37	80.28 ± 4.68
	Average	37.46 ± 4.82	47.33 ± 7.43	47.45 ± 6.78

Table 9: Performance comparison on A&P generation, aggregated over the patient's entire hospital stay. For readers interested in a more detailed breakdown, we refer them to the Fig 10, where we provide per-day performance results for a subset of patients with a length of stay of at least 5 days.

A.6 RAG and direct generation prompt optimization

We optimized the prompts and queries used for RAG systems on all tasks. Both prompts and queries are similar in terms of its instruction, the only difference is that queries for RAG has the "Retrieve" component.

Tables 14,15,16,17,18 present all prompts and queries for MIMIC tasks.

A.7 Impact of context length on consecutive patient data

Figure 10 illustrates the impact of incorporating prior-day context on F1 scores for patient note generation. The x-axis represents consecutive relative days within a patient's hospital stay, where Day 0 corresponds to the first day of admission and serves as the ground truth (i.e., no evaluation is performed on this day). The y-axis shows the average F1 score

Section	Method	CUI	ROUGE-L	BERTScore
Dx	Direct Gen	5.46 ± 10.07	**3.42** ± 4.34	**50.07** ± 10.78
	CoT Prompt	2.80 ± 5.04	2.95 ± 3.48	48.46 ± 11.00
HC	Direct Gen	**8.64** ± 5.98	**12.28** ± 3.13	**62.51** ± 5.95
	CoT Prompt	7.78 ± 5.89	9.98 ± 3.84	59.60 ± 7.32
DI	Direct Gen	9.09 ± 6.90	**12.07** ± 3.90	60.05 ± 9.09
	CoT Prompt	**9.42** ± 8.19	10.83 ± 4.58	**60.52** ± 7.86

Table 10: Mistral performance on the three sections of discharge summarization (Dx: Diagnosis, HC: Hospital course, DI: Discharge Instruction), comparing direct generation approach with the event extraction CoT. Including CUI f-score.

Experiment	Top-K	Chunk Size	Chunk Overlap
Exp 1	10	500	100
Exp 2	20	750	100
Exp 3	50	500	50
Exp 4	20	250	100
Exp 5	50	750	200

Table 11: Hyperparameter configurations for each experiment.

Model	Metric	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Mistral	ROUGE-L	6.14	6.18	5.93	6.10	6.37
	BERTScore	8.41	8.12	8.37	8.35	8.03
	CUI	51.03	50.82	51.12	52.56	51.34
Llama3	ROUGE-L	7.37	7.43	6.95	7.07	6.90
	BERTScore	9.23	9.78	9.56	9.19	9.32
	CUI	55.45	56.90	56.25	55.89	56.67
Qwen	ROUGE-L	6.41	6.97	6.95	6.60	7.11
	BERTScore	9.72	9.62	9.68	9.66	10.09
	CUI	55.76	56.34	55.42	55.90	56.21
Llama2	ROUGE-L	7.56	7.13	7.59	7.42	6.64
	BERTScore	10.22	10.21	9.96	10.16	9.41
	CUI	55.67	54.70	54.37	55.80	53.50

Table 12: Hyperparameter Tuning for Discharge Summaries Results (averaged across sections)

Model	Metric	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Mistral	ROUGE-L	21.07	17.72	15.74	20.30	15.52
	BERTScore	71.42	70.03	68.81	71.45	69.95
	CUI	28.05	21.94	20.63	27.92	19.97
Llama3	ROUGE-L	16.73	17.54	16.74	17.11	16.58
	BERTScore	71.46	71.66	71.02	71.11	71.60
	CUI	22.52	23.88	23.40	23.71	22.50
Qwen	ROUGE-L	19.13	21.13	20.39	18.59	20.62
	BERTScore	70.29	69.34	70.98	70.22	70.03
	CUI	27.28	30.17	28.87	28.04	29.87

Table 13: Hyperparameter Tuning for A&P Generation

across a subset of 10 patients with a length of stay of at least 5 days.

Since Day 0 (Day 1 in the figure) is not evaluated, its F1 score is set to 1.0 for all methods to ensure a clear visual comparison with subsequent days. The results demonstrate that incorporating prior-day context improves performance over time. The Single-Day Context and Multi-Day Context methods achieve substantially higher F1 scores than the No Prior Context method, particularly on Day 2 and Day 3, suggesting that leveraging past information helps generate more accurate and coherent patient notes. However, after Day 3, the performance of the Multi-Day Context method begins to decline, indicating that while longer historical context can be beneficial, it may introduce additional noise or redundant information.

Overall, these findings highlight that considering prior context enhances the accuracy of generated patient notes, with Single-Day Context yielding the highest performance in later days, while Multi-Day Context shows initial improvements but exhibits diminishing returns over time.

A.8 More results for using ground-truth progress notes vs. generated progress notes on A&P Generation

This section presents further results comparing ground-truth progress notes and model-generated progress notes as input for A&P Generation. Our analysis evaluates how using prior ground-truth notes versus LLM-generated notes impacts the overall quality, coherence, and accuracy of the generated A&P sections.

Table 19 presents results on all LLMs direct generation. Overall, there are performance decreases when moving from using generated progress notes as context input. The results highlight key differences in information retention and propagation, where models using ground-truth progress notes tend to maintain better clinical consistency, while those using generated notes may accumulate errors over multiple days, leading to drift and hallucination in longitudinal patient summaries. These findings emphasize the need for robust calibration and error correction mechanisms when relying on LLM-generated progress notes for iterative summarization.

Additionally, we include Figure 11 to show the impact of incorporating prior-day context on the composition of generated patient notes. The figure illustrates how different methods—ranging from no prior-day context (Method -1) to single-day (Method 1) and multi-day context (Method 2)—affect the alignment between generated elements (GEN), ground truth elements (GT), and common elements across multiple admissions. As the number of prior days included increases, we observe changes in the proportion of correctly retained ground truth elements and newly generated con-

What is the patient’s main diagnosis?
The patient’s primary diagnosis is: (Llama2)
Identify the primary reason for the patient’s hospital admission: (Llama3)
Instruct: Given a search query, retrieve relevant passages that answer the query. Query: patient’s primary diagnosis. (Mistral)
The patient has been diagnosed with: (Qwen)

Table 14: Queries used for retrieving the primary diagnosis. Highest performing on models bolded with respective model in parentheses.

Summarize the hospital course for this patient in a concise and accurate way. (Qwen)
The patient’s hospital course included the following:
Provide a brief hospital course, including key events and treatments. (Mistral, Llama3)
Instruct: Given a search query, retrieve relevant passages that answer the query. Query: Brief hospital course.
What were the key events and outcomes during the patient’s hospital stay? (Llama2)

Table 15: Queries used for retrieving the hospital course. Highest performing on models bolded with respective model in parentheses.

tent, highlighting the role of historical context in improving consistency and completeness in patient note generation. Additionally, the inclusion of prior context reduces the number of newly generated elements, indicating a shift away from generating potentially extraneous or less relevant content.

However, the degree of improvement varies across different admissions, suggesting that some cases benefit more from historical context than others, potentially due to differences in case complexity or the structure of prior notes. Furthermore, the differences between generated and ground truth elements become more pronounced as days progress, highlighting the challenge of maintaining consistency in patient notes over time. Overall, the findings suggest that incorporating multi-day context enhances the accuracy and stability of generated patient notes, reducing hallucinated content while preserving clinically relevant information.

A.9 Break-down results behind Figure 5

Recall that Figure 5 aggregates all metrics for LLM direct generation on discharge summarization. To provide comprehensive results analysis, we include Figure 12 for 24h window and 48h window. The trends of performance changing across different modalities are consistent with the Figure 5.

A.10 Details about human evaluation

Last but not least, we present the full survey questions aggregated from existing work in Table 20. We would like to emphasize that this survey has *not* been validated, rather, it selects the criteria after consulting with a physician regarding what they value in LLM generated text and findings from prior work, as cited in the table. These questions allow us to evaluate the accuracy, faithfulness, readability, and clinical relevance of LLM-generated text, providing a structured framework for assessing their strengths and limitations in a clinical setting.

Given the input EHR data, generate discharge instructions for this patient. (All models)
What are the discharge instructions for the patient?
Write a summary of the discharge plan, including medications, follow-up visits, and patient care instructions.
Instruct: Given a search query, retrieve relevant passages that answer the query. Query: discharge instructions.
What follow-up care and medications are recommended for the patient after discharge?

Table 16: Queries used for retrieving the discharge instructions. Highest performing on models bolded with respective model in parentheses.

Given the patient EHR data, write the Assessment section of a clinical progress note. The Assessment should include a brief description of both passive and active diagnoses. Clearly state why the patient is admitted to the hospital and describe the active problem for the day, along with any relevant comorbidities the patient has.
Provide the Assessment section of the patient's progress note, including active and passive diagnoses, admission reasons, the patient's active problems for the day, and relevant comorbidities.
What are the patient's active and passive diagnoses? Why was the patient admitted to the hospital? What are the active medical problems for the day? Include relevant comorbidities.
Retrieve passages that explain the patient's active and passive diagnoses, reasons for admission, active problems for the day, and relevant comorbidities. (Mistral, Qwen)
Instruct: Generate a concise Assessment section for the patient's progress note. Include a summary of active and passive diagnoses, admission reasons, the patient's current active problems, and any comorbidities. (Llama3)

Table 17: Queries used for retrieving the Assessment section of a progress note. Highest performing on models bolded with respective model in parentheses.

Chain-of-Thought Prompt for 48h Discharge Summary Event Extraction

DISCHARGE EVENT EXTRACTION TASK Analyze the following data from the final 48 hours of the hospital stay and identify key clinical events that are most relevant for summarizing the course of treatment and informing discharge planning. {chronology_text} Only include events that reflect:

1. Significant changes in symptoms or status (e.g., improvements, worsening, new findings).
2. Clinically important test results (especially abnormal values that lead to certain treatments).
3. Major treatments or interventions (e.g., medication changes, procedures, escalation/de-escalation).
4. Care team decisions that indicate readiness for discharge or change in care goals.
5. Events linked to the final diagnosis or that inform follow-up care.

Response Format: ### Day X Key Events ### - [Time]: [Description] (Reasoning)

Example Model Output

### Day 1 Key Events ### - ****2159-03-12 08:00****: Stable Vital Signs (BP 120/70 mmHg, HR 88 bpm, RR 18 breaths/min, Temp 98.6°F). This indicates overall stability and readiness for discharge.

Figure 9: Chain-of-Thought prompting template and example output used for discharge summary event extraction based on the final 48 hours of hospitalization.

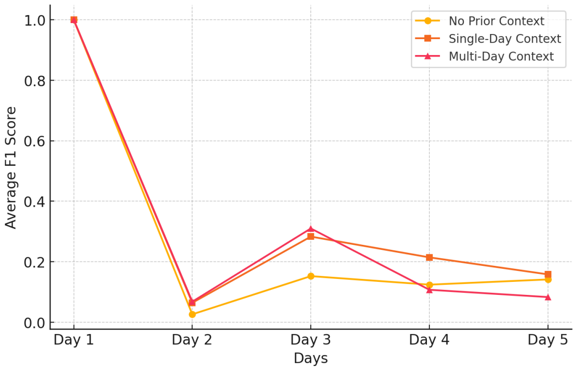

Figure 10: Impact of prior-day context on F1 score across consecutive days

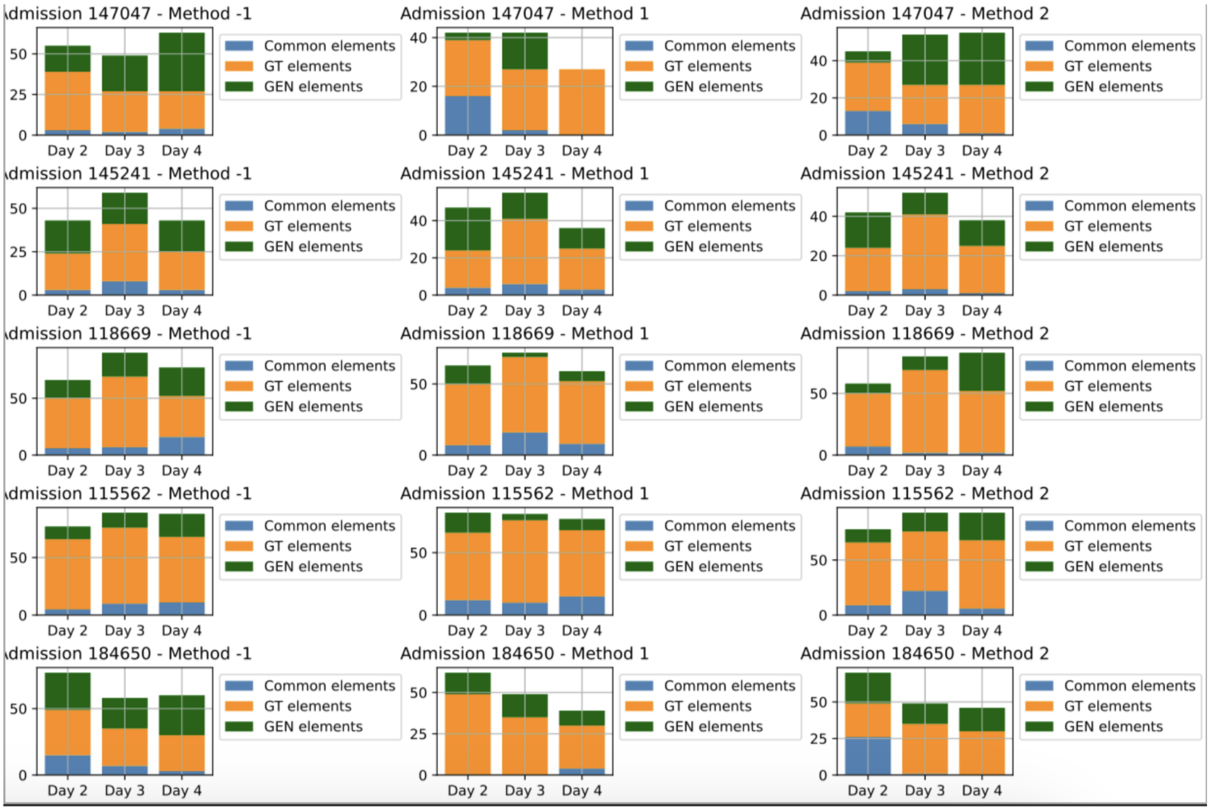

Figure 11: Comparison of generated patient notes across different methods of incorporating prior-day context. Each subplot represents a different patient admission, with bars indicating the composition of generated notes across Days 2, 3, and 4. The stacked bars show the proportion of common elements (overlapping between ground truth and generated notes), GT elements (present in the ground truth but missing in generated notes), and GEN elements (newly generated content not found in the ground truth). Method -1 (no prior-day context), Method 1 (single prior-day context), and Method 2 (multi-day context) demonstrate how historical context influences the balance between accurate retention and newly introduced information.

Given the patient EHR data, write the Plan section of a clinical progress note. The Plan should be organized into multiple subsections, each corresponding to a specific medical problem. Provide a detailed treatment plan for each problem, outlining proposed or ongoing interventions, medications, and care strategies.
Generate the Plan section of the patient's progress note. Provide detailed treatment plans for specific medical problems, including proposed or ongoing interventions, medications, and care strategies.
What are the proposed treatment plans for the patient's active medical problems? Include details on interventions, medication regimens, and care strategies.
Retrieve passages that outline treatment strategies for medical problems, including medications, interventions, and care strategies. **(Mistral, Qwen)**
Instruct: Write the Plan section of the progress note. Organize it into subsections for each medical problem. Provide detailed plans for treatments, interventions, medications, and care strategies. (Llama3)

Table 18: Queries used for retrieving the Plan section of a progress note. Highest performing on models bolded with respective model in parentheses.

			GT			**GEN**	
Model	**Metric**	**No Prior**	**Single-Day**	**Multi-Day**	**No Prior**	**Single-Day**	**Multi-Day**
Mistral	CUI	21.60±6.47	34.06±12.15	30.55±13.48	21.60±6.47	22.30±16.13	22.82±16.47
	ROUGE-L	17.70±3.79	26.11±11.68	23.32±13.05	17.70±3.79	17.57±13.47	17.90±13.59
	BERTScore	68.16±6.91	71.63±10.86	72.41±9.47	68.16±6.91	70.51±11.60	71.50±9.66
	Average	35.82	43.93	42.09	35.82	36.79	37.41
Llama3	CUI	21.83±6.51	45.35 ± 10.45	43.55±12.00	21.83±6.51	40.89±13.61	40.50±13.29
	ROUGE-L	16.60±3.39	33.16±10.68	32.42±12.14	16.90±3.85	28.86±12.8	29.34±13.30
	BERTScore	72.01±6.04	80.10±8.93	80.42±8.81	72.33±5.48	79.68±10.07	80.48±9.56
	Average	36.81	52.87	52.13	37.05	49.81	50.11
Qwen	CUI	21.84±5.68	34.27±7.18	34.32±6.65	21.84±5.68	30.10±6.70	29.55±6.50
	ROUGE-L	15.35±3.10	77.50±5.37	21.85±4.99	15.18±3.00	18.21±4.28	17.55±4.07
	BERTScore	72.57±4.92	20.69±4.37	77.88±4.86	72.97±4.83	78.39±4.11	77.61±4.52
	Average	36.87	44.15	44.68	37.12	42.23	41.57
DeepSeek	CUI	22.58±6.63	34.45±9.26	34.15±10.00	22.58±6.63	27.33±7.69	27.93±7.46
	ROUGE-L	16.16±3.15	21.84±5.65	21.07±5.42	16.16±3.15	17.37±3.86	16.88±3.88
	BERTScore	74.97±3.66	78.29±4.61	78.35±5.12	74.97±3.66	77.90±4.36	77.93±4.79
	Average	37.90	44.86	44.52	37.90	40.87	40.91

Table 19: Comparison of ground-truth (GT, same results we report in main text) and generated (GEN) settings on assessment and plan generation (results of direct generation approach shown)

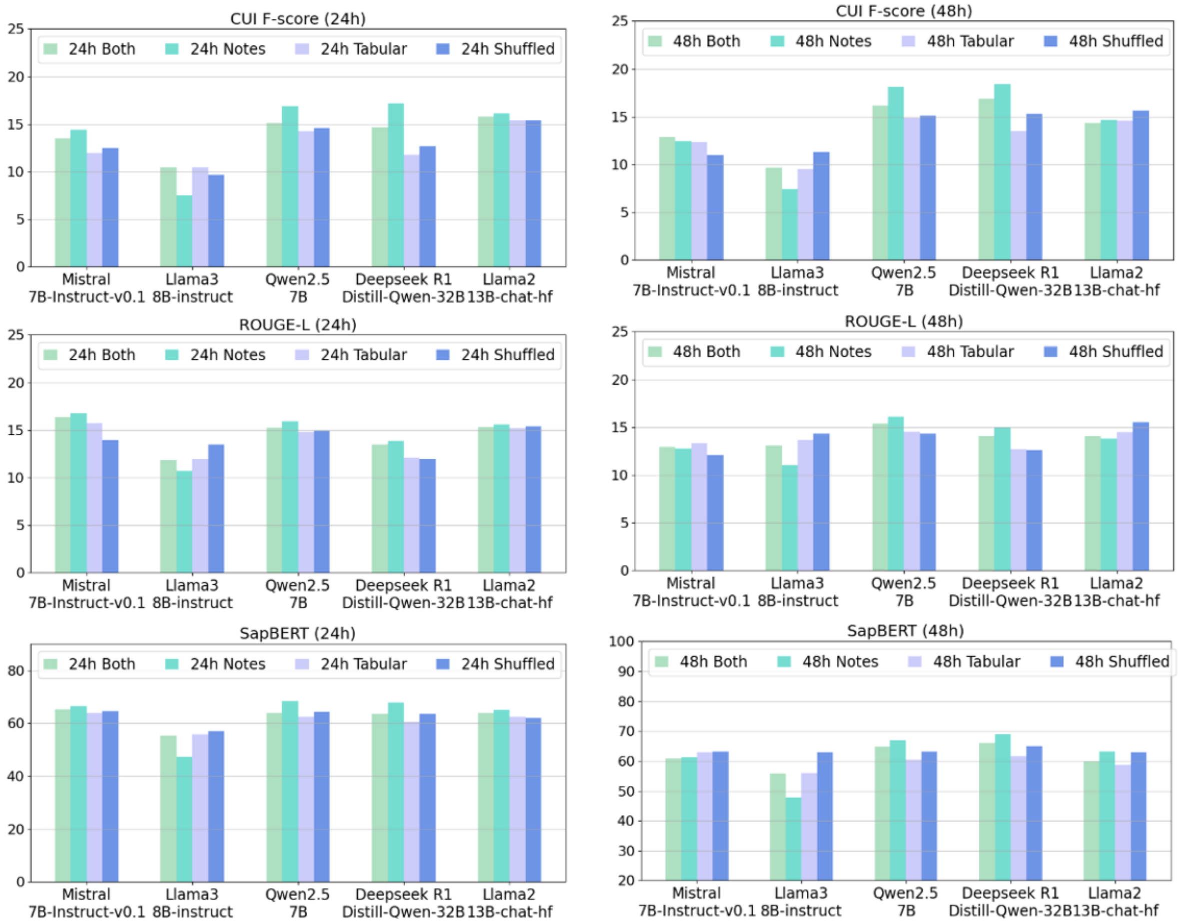

Figure 12: Metric breakdown across modalities for the 24 hour time window (on direct generation discharge summarization). SapBERT = BERTScore with SapBERT emebddings.

Physician Evaluation of LLM-Generated Clinical Summaries

Thank you for participating in this evaluation. Please assess the LLM-generated text based on the following criteria using the 5-point Likert scale provided for each question.

Scale Definitions (from 1-5):

1 Strongly Disagree; 2 Disagree; 3 Neutral; 4 Agree; 5 Strongly Agree

Category	Evaluation Question	Score (1-5)
Overall Accuracy (Xu et al., 2024a; Singhal et al., 2023; Ben Abacha et al., 2023; Croxford et al., 2025; Johri et al., 2025)	How well does the generated text align with the actual clinical data? The summary is factually correct and accurately represents the original data. No major distortions or misinterpretations of key clinical facts.	
Hallucination (Faithfulness) (Ben Abacha et al., 2023; Singhal et al., 2023; Aljamaan et al., 2024; Johri et al., 2025)	To what extent does the LLM generate information faithfully? The generated text does not introduce any fabricated, misleading, or incorrect information. All statements in the summary can be traced back to the original document.	
Omission (Croxford et al., 2025; Ben Abacha et al., 2023)	Did the model include all important clinical details? The generated summary includes all clinically important details. No missing critical pieces of information relevant to patient care.	
Readability (Xu et al., 2024a)	How easy is the generated text to read and comprehend? The text is well-structured, clear, and easy to understand. Uses appropriate medical terminology without excessive complexity.	
Clinical Relevance (Usefulness) (Singhal et al., 2023; Johri et al., 2025)	How useful is the content for clinical decision-making? The information provided is highly relevant to the clinical task. Avoids unnecessary or unrelated details.	
Specificity (Level of Detail) (Williams et al., 2024)	Does the summary maintain an appropriate level of detail? The text balances between a high-level summary and necessary details. Avoids being overly vague or excessively detailed.	

Table 20: Physician evaluation survey