

CliniCIRCA: A Modular LLM Framework for Constructing Longitudinal Mental Health Patient Journeys from Raw EHR Narratives

Aiwei Ivy Zhang¹ Nimra Ishfaq² Mohit Chandra¹ Santiago Alvarez Lesmes³
Adam Coscia¹ Khatiya Chelidze Moon³ Xiaohan Ding¹ Munmun De Choudhury¹

¹Georgia Institute of Technology ²University of Texas at Austin ³Northwell Health

Abstract

In mental health care, reasoning over patient journeys is a key task for clinicians. Yet these journeys, encompassing a longitudinal progression of biological, psychological, and social events, are often spread across disparate unstructured text narratives, making temporal recovery challenging. We present **CliniCIRCA**, a multi-stage LLM framework for Calendar-anchored, Imprecision-aware **R**econstruction of **C**linical **A**nnals. To our knowledge, CliniCIRCA is the first to temporally classify clinical events across unstructured discharge summaries without event-level timestamps. From 14,882 MIMIC-III mental health admissions, we first construct a benchmark of 52 discharge summaries on which CliniCIRCA produces 15,891 temporally tagged events. After correcting 629 errors based on a clinician-in-the-loop evaluation, we produce verified gold-standard labels. Finally, the corrected timelines drive a temporally grounded summarization stage that compresses each source 1.52 times into a date-grouped chronological record. We then scale the framework to generate 1,000 silver-standard timelines and evaluate them as training data. Compared with zero- and few-shot prompting, instruction tuning generally improves five open-weight models on event extraction, temporal tagging, and summarization across silver and clinician-verified evaluations.

1 Introduction

Every patient follows a clinical journey: a longitudinal progression of symptoms, diagnoses, treatments, life events, and responses to care that demonstrate how their health has evolved over time (Kraljevic et al., 2024; Ruan et al., 2019). Clinicians draw on this progression to interpret a patient’s current presentation and guide treatment.

Unfortunately, these journeys are often not documented as explicit timelines (Loftus et al., 2024; Linhares et al., 2023; Olex and McInnes, 2021).

For example, discharge summaries are a rich source of longitudinal patient information. They synthesize prior history, hospital course, and clinical outcomes in free-form text, where events may be repeated, narrated nonsequentially, or linked only through relative temporal expressions and surrounding context (K et al., 2025; Seinen et al., 2025; Lee et al., 2024; Kim et al., 2024).

Recovering temporal structure is thus a central challenge in clinical NLP for modeling patient journeys from clinical text (Amirahmadi et al., 2025; Makarov et al., 2025). Patient journeys are often spread across disparate text sources, limiting the use of these narratives by computational systems that require structured longitudinal representations (Chaturvedi, 2024; Moharasan and Ho, 2019). As a result, information fragmentation shifts the burden of reconstruction to clinicians and computational systems. Clinicians must piece together the patient’s course under time and information constraints, while models must infer temporal relations that remain implicit (Linhares et al., 2023; Gao et al., 2022; Olex and McInnes, 2021).

Prior work treated event reconstruction as a set of separable subtasks, which makes temporal alignment challenging. For example, clinical summarization compresses lengthy records into coherent accounts, but often under-preserved temporality and rarely represents complex or uncertain event chronology explicitly (Kruse et al., 2025; Cui et al., 2025; Croxford et al., 2025). Temporal information extraction preserves finer structure by extracting time expressions, events, and temporal relations, but most systems are built around predefined annotation schemas and pairwise relation classification rather than recovery of an open-ended patient trajectory (Chaturvedi et al., 2025; Gumiel et al., 2022; Alfattni et al., 2020).

Recent LLM timeline generation comes closer to recovering full patient trajectories (Kumar et al., 2026; Noroozizadeh et al., 2025; Wang and Weiss,

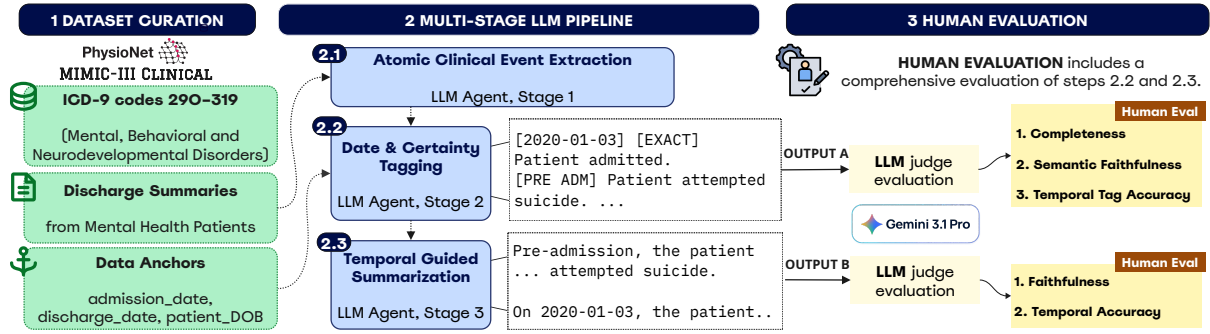


Figure 1: Methodology Overview

2025; Wang et al., 2024), but often relies on structured EHR timestamps, preselected note fragments, or curated case reports, and commonly represents time using a single relative offset. What remains missing is a unified, text-only reconstruction formulation that preserves event breadth, calendar timing, temporal uncertainty, and patient-level coherence.

We specifically focus on bridging the gap in mental health care, where a patient’s course cannot be understood only from diagnoses and procedures alone. Engel’s framework emphasizes that illness emerges through the interaction of biological, psychological, and social factors (Engel, 1977); in mental health, these factors are often expressed through medication use, trauma, social history, changes in daily functioning, etc (Saha et al., 2026). Such information is part of the trajectory rather than peripheral context, yet the combination of events differs across patients. Mental health records therefore expose the limitations of fixed schemas: a narrow schema would omit events that may define an individual cause, while an overly exhaustive one quickly becomes impractical.

Hence, we formulate patient-journey recovery as **open-vocabulary, calendar-anchored, and uncertainty-aware temporal reconstruction from raw clinical narratives**. Specifically, *journeys are reconstructed within a single discharge summary corresponding to one hospital admission, as a step toward multi-encounter longitudinal modeling*. We introduce **CliniCIRCA** (Figure 1), a three-stage LLM framework to convert discharge summaries to coherent temporal clinical trajectories. CliniCIRCA produces two linked representations: Output A, an inspectable event-level timeline for computational use and auditing, and Output B, a clinician-readable account derived from the same temporal substrate. Our experiments demonstrate that patient journeys produced by a high-capacity model (Section 6) can provide effective supervision

for substantially smaller models, particularly for atomic clinical event extraction, while temporal reconstruction remains reasoning-intensive. By transforming fragmented narratives into linked structured and readable representations, CliniCIRCA could support longitudinal chart review, clinical handoffs, and follow-up care while preserving an auditable connection to the underlying events. We have provided code in a repository here¹.

2 Dataset

We use MIMIC-III v1.4 (Johnson et al., 2016, 2015), a publicly available, de-identified critical-care database from the Beth Israel Deaconess Medical Center (2001–2012) in English. CliniCIRCA receives unprocessed discharge summaries without chunking, section parsing, or normalization. Structured EHR fields are used to select the cohort and provide three temporal anchors: (1) date of birth, (2) admission date, and (3) discharge date. We define the cohort at the admission level: an admission is retained if it contains at least one psychiatric diagnosis from the ICD-9 Mental Disorders chapter (codes 290–319) and can be paired with a discharge summary. This produces 14,882 summaries from 12,273 patients, each summary treated as a separate reconstruction instance. Because admissions are flagged by the presence of any Mental Disorders diagnosis code rather than by principal diagnosis, the cohort includes both primary psychiatric admissions and medical/surgical admissions with psychiatric comorbidity. More details can be found in Appendix A.

3 The CliniCIRCA Framework

CliniCIRCA is a sequential framework comprising three LLM stages (Figure 1), each receiving

¹<https://anonymous.4open.science/r/CliniCIRCA-E5FF/>

the output of the previous stage and having a single main task. **Stage 1** reads the raw discharge summary (obtained from the MIMIC-III Dataset) without preprocessing and extracts a high-recall list of atomic clinical events. **Stage 2** resolves the timing of each event, tagging it with an ISO date and a certainty label anchored against temporal references; its output is **Output A**, a timeline of (ISO date and certainty tag, atomic event) pairs. **Stage 3** aggregates these tagged events by date into **Output B**, a date-grouped chronological summary. Decomposing the task in this way allows each stage to be prompted, evaluated, and improved independently, and confines each temporal decision to the stage best positioned to make it.

4 Model and Agent Evaluation Setup

4.1 Candidate Pool

For CliniciCIRCA, we select 11 LLM candidates based on reported performance across general-purpose and clinical benchmarks (Yu et al., 2025; Sarvari and Al-fagih, 2025; Shi et al., 2024). For proprietary models, we evaluate Gemini 2.5 Pro and 2.5 Flash (Comanici et al., 2025); Claude Sonnet 4.6 (Anthropic, 2026) and Haiku 4.5 (Anthropic, 2025). For open-weight models, we evaluate Llama 3.3 70B (Meta AI, 2024), Qwen 3.6 35B-A3B (Qwen Team, 2026), Gemma 4 26B-A4B (Gemma Team, 2026), Mistral Small 3.2 24B-A4B (Mistral AI, 2025), Phi-4 Mini 3.8B (Abouelenin et al., 2025), and medical-domain LLMs - MedGemma 27B (Sellersgren et al., 2026) and Meditron 3 Qwen 2.5 7B (OpenMeditron, 2025). Model specifications, generation settings, and computational details are provided in Appendix B.

4.2 Implementation Setup

We construct a benchmark set of 52 MIMIC-III discharge summaries, averaging 1,774 words each (range: 437–3,583; Appendix A.2). Of the 52 cases, 29 had a principal psychiatric diagnosis; the remaining 23 were medical or surgical admissions with psychiatric comorbidity. For the benchmark, every candidate is run through Stages 1-2, with no mixing of models across stages. The size is determined by a power analysis (Appendix C) for a pairwise χ^2 comparison (large effect, $\alpha = 0.05$, 95% power), yielding $N = 52$ needed to reach statistical significance (Ki et al., 2025; Card et al., 2020; Dror et al., 2018).

Each candidate is run through Stages 1-2 as

a fixed pipeline configuration, producing 572 model–document outputs evaluated using the clinician-authored rubric described in Section 5.1. We retain the selected model for Stage 3 rather than optimizing a separate model for each stage. This avoids a combinatorial search over model combinations, reveals how one model’s capabilities propagate through the CliniciCIRCA framework, and simplifies reproducibility and deployment governance (Umeton et al., 2024; de Hond et al., 2024; Meskó and Topol, 2023). We select based on Stage 2 because the summary stage depends on the accuracy of its event-level temporal input. Following prior health-domain work that validates LLM judges against expert-vetted human labels (Mittal et al., 2026), two human annotators and two clinicians, all with relevant annotation experience and high proficiency in English, form a feedback loop to evaluate the results. We establish human-evaluator agreement on both Stage-2 and 3 outputs.

5 Constructing Patient Timelines from Discharge Summaries

5.1 Stage 1 & 2: From Raw Discharge Summary to Tagged Timeline

Faithful clinical timeline recovery requires both comprehensive event extraction and accurate temporal anchoring, yet extraction errors can propagate through temporal reasoning and timeline construction (Gumiel et al., 2022; Alfattni et al., 2020). We therefore decompose the task into Stage 1 for event extraction and Stage 2 for temporal anchoring, allowing each component to be independently specified and evaluated. Full generation prompts can be found in Appendix D Tables A5 - A6.

Inclusive extraction and anchor-based tagging. Stage 1 reads the raw discharge summary, applying no chunking, section parsing, or normalization, and extracts events inclusively and faithfully. Accordingly, Stage 2 then resolves each event against the admission date, discharge date, and date of birth, assigning exactly one of four temporal categories: 1) [DATE/RANGE] [EXACT]: a date written in the event text itself. 2) [DATE/RANGE] [APPROX]: no written date, but the event resolves to an anchor or a relative offset from one. 3) [PRE ADM]: framed as prior to or continuous through this admission (history, “status post,” demographics, undated chronic states), and 4) [INDETERMINATE]: a guardrail used only when none of the above applies and the timing genuinely cannot be determined.

Model	Dimension A	Dimension B	Dimension C	Avg. ABC	MED-P	MED-R	MED-F1
Gemini 2.5 Pro	5.00	5.00	4.83	4.94	0.637	0.812	0.710
Claude Haiku 4.5	5.00	<u>4.98</u>	<u>4.81</u>	<u>4.93</u>	0.610	0.753	0.672
Claude Sonnet 4.6	<u>4.98</u>	5.00	4.77	4.92	0.614	0.735	0.666
Gemini 2.5 Flash	<u>4.98</u>	5.00	4.50	4.83	0.617	<u>0.762</u>	0.679
Qwen 3.6 35B-A3B	5.00	5.00	4.23	4.74	0.630	0.758	<u>0.685</u>
Gemma 4 26B-A4B	4.88	4.96	4.27	4.71	0.636	0.713	0.669
Mistral Small 3.2 24B-A4B	4.96	<u>4.98</u>	3.88	4.61	0.636	0.740	0.681
Llama 3.3 70B	4.77	4.90	3.44	4.37	0.626	0.673	0.644
MedGemma 27B	<u>4.98</u>	<u>4.98</u>	2.67	4.21	0.638	0.729	0.678
Meditron 3 (Qwen 2.5 7B)	3.71	4.73	2.00	3.48	<u>0.680</u>	0.285	0.361
Phi-4 Mini 3.8B	3.90	4.17	2.19	3.42	0.685	0.461	0.525

Table 1: Stage-2 model-selection results on the 52-sample benchmark. The highlighted row indicates the model selected for the final pipeline. Results are averaged over 52 samples. Dimension A: Completeness; Dimension B: Semantic Faithfulness; and Dimension C: Temporal Tag Accuracy. Dimensions A to C are Likert-scale averages reported to two decimal places. The best-performing value for each metric is shown in **bold**, and the second-best value is underlined. Tied values receive the same formatting.

As specified above, to maximize precision, only explicitly written dates receive [EXACT]; any inferred or relative date resolves to [APPROX] at the coarsest defensible granularity (e.g., year-only when only a year is recoverable). Stage 2’s output is a list of temporally classified clinical events.

Operationalizing the notion of “faithfulness”.

Because no reference timeline exists against which these tuples could be scored, a “good” timeline must itself be operationalized. We define three clinician-informed criteria for evaluating the Stage-2 output (full definitions in Appendix E).

A) Completeness: *whether all clinically relevant events in the note are captured.*

B) Semantic faithfulness: *whether each extracted event preserves the meaning of the source text.*

C) Temporal classification accuracy: *whether the assigned date and tag are correct.*

Calibrating the automated judge. Following established practice in LLM evaluation to use a strong later-generation model (Zheng et al., 2023), we select Gemini 3.1 Pro as the LLM evaluator, chosen to be at least as capable as any candidate it scores. Prior to selection, the evaluator is calibrated against expert human scoring. Only the calibrated LLM evaluator is used in the selection and scaling.

Selecting the agent model. Following the agent selection mechanism as described in Section 4.2, Gemini 2.5 Pro is selected as the strongest configuration, ranking among the top candidates on every dimension and achieving 5.00 on Completeness, 5.00 on Semantic Faithfulness, and 4.83 on Temporal Tag Accuracy in Table 1. As a judge-

independent check, we also report MEDCON (Wai Yim et al., 2023), the F1 overlap of UMLS medical concepts between each model’s extracted events and the source note; once again, Gemini 2.5 Pro attains the highest MEDCON F1 (Table 1), corroborating its judge-scored ranking. We select based on Stage 2 alone and then freeze Gemini 2.5 Pro as the only agent advancing to Stage 3.

5.2 Validating the Stage 1-2 Timeline

We validate the output of Stage 1-2 through a human review with two psychiatric clinicians during 6 iterative annotation sessions. The human review process was used both to evaluate the automated judge and to correct all confirmed errors, providing a clinician-verified gold reference dataset ($N = 52$). Evaluation was conducted on an internally hosted, custom web-based secure platform, demonstrated in Appendix F.

We compare the calibrated Gemini 3.1 Pro evaluator with the clinician-verified gold annotation (see Appendix E for complete rubrics). Overall agreement is high, with Cohen’s $\kappa = 0.655$. However, because only 3.96% of events were labeled as errors, class imbalance likely depresses κ (Gwet, 2008). This is reflected in a Gwet’s AC1 of 0.972 and negative agreement of 0.986, indicating strong agreement on events judged correct. Agreement is lower for the rare error class, with positive agreement of 0.669.

5.3 Stage 3: From Timeline to Temporally-Guided Summary

Next, Stage 3 tests if the verified event timeline can be rendered as a faithful, temporally grounded

Metric	Value
Mean reduction (avg. per-note %)	31.53%
Median per-note reduction	33.76%
Corpus-level reduction (total words)	34.01%
Compression factor	1.52×
Mean words reduced per admission	603
Median words reduced	422

Table 2: Reading-burden reduction from Stage 3 summarization over the 52-note benchmark. Per-note percentages are averaged across notes; corpus-level reduction pools word counts across all notes.

summary without re-reading the discharge note or reconstructing its chronology. Two clinicians defined a specification for the mental-health cohort that retains demographics and social history, reason for admission, key in-stay events, medication changes, discharge, and follow-up. Each event is grouped under its Stage 2 temporal label.

Stage 3 is evaluated for faithfulness to the verified timeline and temporal accuracy using a separately calibrated Gemini 3.1 Pro evaluator validated against clinician judgments. Generation prompts and rubrics appear in Appendices H - I. Across the benchmark, Stage 3 reduces reading burden by roughly a third (31.5% mean and 33.8% median per-note; 34.0% corpus-wide), a stable 1.52 times compression that removes ~ 600 words per admission (Table 2).

5.4 Validating the Stage 3 Summary

We compare the judge’s error counts against human adjudication across 52 summaries (Table A18). The judge flags 262 errors, of which human annotators uphold 39: a 6.7-fold inflation, averaging 4.3 spurious errors per summary. Agreement is substantial on whether a summary contains any error ($\kappa = 0.77$, positive agreement = 0.84) and on the relative ordering of summaries by error count (Spearman $\rho = 0.77$). With clinician guidance, we develop a two-dimensional rubric (Tables A13-A14) to validate if generated summaries: (A) faithfully retained relevant mental-health and social events without distortion and (B) placed them accurately in time. See Appendix K for full metrics and results, and Appendix L for the internally hosted, custom web-based evaluation platform.

5.5 Qualitative Analysis

We qualitatively investigate what this reconstructed timeline helps a reader (e.g., a clinician) understand. To examine the clinical content captured by the

timelines, we analyze three clinician-verified gold cases using two World Health Organization frameworks: eight categories of mental disorders and associated symptoms (World Health Organization, 2025a), and three levels of risk factors: Individual, Family & Community, and Structural (World Health Organization, 2025b). We label a risk factor as co-documented rather than causal unless the record explicitly links it to the disorder. Across three exemplar cases, temporal reconstruction distinguishes long-standing risk factors from acute in-stay changes and diagnostic reassessment, while preserving whether their relationships are causal, co-documented, or unresolved rather than collapsing them into a single causal narrative. Full case analyses are provided in Appendix M.

6 Evaluation with Open-Weight Models

We evaluate whether CliniciCIRCA’s intermediate outputs can serve as reusable supervision under three training and evaluation settings.

SILVER $\rightarrow$ SILVER measures how well a model fits pipeline-generated supervision. We sample 1,000 admissions disjoint from the 52 gold examples, label them with the frozen Stage 1–2 pipeline, and split them into 750 training, 100 validation, and 150 held-out test documents. This Stage 1 to 3 generation processed 61.84M input and output tokens in total, averaging 61,836 tokens per admission across the complete framework (Appendix A.3).

SILVER $\rightarrow$ GOLD measures whether that supervision transfers to human-verified annotations: models trained on the same 750 silver documents are evaluated on all 52 gold examples.

GOLD measures what a small clinician-verified set alone can support. Because 52 examples cannot sustain a stable held-out split, we run four-fold cross-validation and pool the out-of-fold predictions before computing metrics and bootstrap intervals. Throughout, the gold examples are excluded from silver labeling, silver-trained adapters, and few-shot exemplar selection to prevent unintentional leakage. Under the GOLD setting, each document is evaluated only by the fold-specific model for which it was held out.

6.1 Experimental Setup

Models. We evaluate five locally deployable open-weight models spanning dense and mixture-of-experts architectures, 7B–35B parameter scales, and general-purpose versus clinical specializa-

Model	SILVER→SILVER			SILVER→GOLD			GOLD		
	Zero	Few	IT	Zero	Few	IT	Zero	Few	IT
Gemma4-31B	.640 _{.011}	.665 _{.016}	.840 _{.008}	.639 _{.012}	.659 _{.017}	.827 _{.014}	—	.727 _{.013}	.628 _{.016}
Qwen3.6-35B-A3B	.536 _{.014}	.642 _{.013}	.839 _{.009}	.535 _{.021}	.647 _{.017}	.834 _{.012}	—	.732 _{.012}	.550 _{.026}
Qwen3-30B-A3B	.420 _{.027}	.547 _{.025}	.745 _{.028}	.454 _{.028}	.561 _{.020}	.795 _{.013}	—	.279 _{.045}	.606 _{.033}
Mistral-24B	.309 _{.021}	.565 _{.016}	.848 _{.008}	.284 _{.021}	.545 _{.032}	.847 _{.009}	—	.676 _{.015}	.715 _{.028}
Meditron-7B	.040 _{.006}	.131 _{.013}	.744 _{.021}	.032 _{.007}	.120 _{.015}	.752 _{.021}	—	.034 _{.004}	.331 _{.024}

Table 3: Task A measured by embedding-based RM-F1@0.85. SILVER→SILVER trains and evaluates on the silver split; SILVER→GOLD trains on silver and evaluates on all 52 clinician-verified examples; GOLD uses four-fold cross-validation on the gold set alone. Results are point estimates $\pm$ bootstrap SD.

Model	SILVER→SILVER			SILVER→GOLD			GOLD		
	Zero	Few	IT	Zero	Few	IT	Zero	Few	IT
Gemma4-31B	.738 _{.019}	.731 _{.019}	.767 _{.024}	.799 _{.036}	.804 _{.043}	.749 _{.029}	—	.668 _{.029}	.721 _{.035}
Qwen3.6-35B-A3B	.635 _{.021}	.687 _{.020}	.742 _{.019}	.567 _{.020}	.715 _{.025}	.731 _{.022}	—	.602 _{.023}	.633 _{.019}
Qwen3-30B-A3B	.426 _{.017}	.609 _{.021}	.695 _{.016}	.452 _{.019}	.628 _{.031}	.658 _{.011}	—	.231 _{.038}	.425 _{.020}
Mistral-24B	.573 _{.013}	.638 _{.024}	.733 _{.025}	.567 _{.040}	.615 _{.023}	.734 _{.018}	—	.620 _{.012}	.680 _{.007}
Meditron-7B	.266 _{.016}	.004 _{.003}	.401 _{.035}	.256 _{.018}	.004 _{.004}	.666 _{.048}	—	.005 _{.004}	.359 _{.022}

Table 4: Task B temporal tagging, scored by macro-averaged F1 over the four temporal tags with the atomic events supplied as oracle input. Results are point estimates $\pm$ bootstrap SD.

Model	SILVER→SILVER			SILVER→GOLD			GOLD		
	Zero	Few	IT	Zero	Few	IT	Zero	Few	IT
Gemma4-31B	.671 _{.011}	.647 _{.011}	.792 _{.010}	.636 _{.015}	.616 _{.019}	.749 _{.017}	—	.570 _{.026}	.609 _{.028}
Qwen3.6-35B-A3B	.538 _{.011}	.584 _{.011}	.779 _{.010}	.490 _{.016}	.540 _{.017}	.754 _{.015}	—	.497 _{.029}	.516 _{.032}
Qwen3-30B-A3B	.504 _{.012}	.532 _{.012}	.705 _{.014}	.474 _{.015}	.522 _{.015}	.674 _{.021}	—	.108 _{.032}	.214 _{.030}
Mistral-24B	.513 _{.012}	.538 _{.012}	.770 _{.011}	.450 _{.019}	.486 _{.017}	.738 _{.017}	—	.484 _{.019}	.616 _{.025}
Meditron-7B	.361 _{.016}	.328 _{.023}	.690 _{.018}	.293 _{.021}	.244 _{.031}	.663 _{.021}	—	.101 _{.018}	.159 _{.020}

Table 5: Task C timeline summarization measured by ROUGE-L. Models receive temporally tagged events and the source discharge summary as oracle input. Results are point estimates $\pm$ bootstrap SD.

tion: Gemma4-31B (Gemma Team, 2026), Mistral-24B (Mistral AI, 2025), Qwen3.6-35B-A3B (Qwen Team, 2026), Qwen3-30B-A3B (Qwen Team, 2025), and Meditron-7B (OpenMeditron, 2025). Qwen3-30B-A3B and Gemma4-31B were not evaluated during Stage 2 model selection, allowing us to test whether observed performance extends beyond the models previously selected.

Prompting and tuning. For each task, we compare three settings under identical task prompts: zero-shot prompting, few-shot prompting with three training exemplars, and LoRA-based instruction tuning with one adapter per model and task. Instruction tuning is intended to align models to the expected output format and decision rules rather than to inject new clinical knowledge.

Metrics. We evaluate Task A using relaxed-match F1 with a cosine similarity threshold of 0.85 (RM-F1@0.85) based on prior research (Sharif et al., 2025; Zhang et al., 2019). Task B uses macro-

F1 over four temporal categories, with a prediction considered correct when both the temporal label and resolved ISO date or range match the reference. Task C is evaluated using ROUGE-L; ROUGE-1 and ROUGE-2 are reported in Appendix N.

Statistical comparison. We report point estimates on the full evaluation set with standard deviations from 2,000 document-level bootstrap resamples. For comparisons on the same evaluation set, resamples are paired across systems, and differences are assessed using 95% confidence intervals for the paired score difference. We describe a difference as reliable when its confidence interval excludes zero; otherwise, we make no claim that one system outperforms the other.

6.2 Results

We examine if silver supervision improves open-weight models, transfers to clinician-verified annotations, and provides greater utility than the gold set

alone. Note that zero-shot prompting requires no training fold, so its GOLD-condition performance is identical to the zero-shot SILVER→GOLD condition; we therefore omit the redundant GOLD zero-shot column in Tables 3-5. Unless otherwise noted, comparisons below describe point-estimate rankings; we flag a difference as statistically reliable only where the paired bootstrap confidence interval excludes zero, and treat other orderings as suggestive rather than established.

Task A: Atomic event extraction. Under SILVER→GOLD, instruction tuning yields higher RM-F1@0.85 scores than zero-shot and few-shot prompting for all five models (Table 3). Mistral-24B achieves the highest point estimate at 0.847, compared with 0.284 zero-shot and 0.545 few-shot. Silver-trained models also retain similar point estimates when evaluation shifts from the silver test set to clinician-verified annotations. For example, Mistral-24B scores 0.848 under SILVER→SILVER and 0.847 under SILVER→GOLD. These results show that benefits of silver supervision extend beyond evaluation against the generated labels.

Task B: Temporal tagging. Temporal tagging shows a less uniform effect of instruction tuning (Table 4). Under SILVER→GOLD, instruction tuning yields higher scores than both prompting baselines for four of the five models. Mistral-24B and Qwen3.6-35B-A3B reach 0.734 and 0.731, respectively. Gemma4-31B is the exception, achieving its highest score with few-shot prompting (0.804), compared with 0.799 zero-shot and 0.749 after instruction tuning.

Task C: Timeline summarization. Under SILVER→GOLD, instruction tuning yields higher ROUGE-L scores than both prompting baselines for all five models (Table 5). Qwen3.6-35B-A3B achieves the highest point estimate at 0.754, followed by Gemma4-31B at 0.749 and Mistral-24B at 0.738. Meditron-7B shows the largest increase, rising from 0.293 zero-shot and 0.244 few-shot to 0.663 after instruction tuning.

Silver supervision is more effective than gold-only tuning at the current scale. Across all three tasks and all five models, instruction tuning on the silver dataset yields higher gold-set point estimates than training on the gold folds alone. For Task A, Mistral-24B achieves 0.847 under SILVER→GOLD, compared with 0.715 under

GOLD; Qwen3-30B-A3B scores 0.795 and 0.606, respectively. The same pattern appears in Task B.

The gold-only results should not be interpreted as evidence that the silver annotations are of higher quality. Rather, they show that at the evaluated scale, broader coverage and larger volume of the silver dataset provide sufficient supervision for downstream models to learn the target tasks. Gold annotations remain essential as an independent benchmark for assessing whether models trained on silver supervision generalize to human-verified outputs.

7 Related Work

7.1 Temporal Annotation in Clinical NLP

Clinical temporal NLP has traditionally focused on extracting typed events, time expressions, and pairwise temporal relations in benchmarks such as i2b2, THYME, and n2c2 (Styler et al., 2014; Sun et al., 2013; Uzuner et al., 2011). Later work extended this paradigm toward absolute event timing and temporal intervals (Leeuwenberg and Moens, 2020; Frattallone-Llado et al., 2024). These methods established the foundations of clinical temporal reasoning, but generally assume predefined event types or annotated mentions (Wang et al., 2024). ClinICIRCA instead reconstructs an open-vocabulary, patient-level timeline directly from narrative text, grounding events to calendar dates or ranges while preserving temporal uncertainty.

7.2 Constructing Event Timelines & Summaries from Clinical Narratives

Recent work has used LLMs to generate event time sequences from clinical narratives through either single-pass extraction or multi-stage reconstruction (Kumar et al., 2026; Noroozizadeh et al., 2025; Wang and Weiss, 2025). Existing approaches make this task tractable by imposing different forms of structure. MedTimeline focuses on a predefined schema of chemotherapy events and does not model temporally vague events (Wang et al., 2024). Frattallone-Llado et al. (2024) and Kumar et al. (2026) rely on structured EHR data to improve or calibrate event timing. MIMIC-IV-Ext-22MCTS applies chunking and retrieval to discharge summaries before generating events with relative timestamps (Wang et al., 2025a). In contrast, Noroozizadeh et al. (2025) construct timelines from published case reports, which are typically more organized and curated than routine clinical narratives. Together, these studies demonstrate the feasibility

and downstream value of narrative timeline generation, but they focus on settings that are narrower, more structured, or more curated than ours.

Clinical summarization addresses a complementary goal: condensing lengthy records into coherent accounts for review and decision-making (Luo et al., 2023; Lee et al., 2024; Croxford et al., 2025). Most summarization systems generate directly from the source note, combining event selection, temporal reasoning, and narrative generation in a single step that makes the underlying chronology difficult to inspect. CliniciCIRCA instead decouples these decisions: it first constructs an event-level, uncertainty-aware timeline and then summarizes from that representation. The summary is therefore not an independent compression of the original note, but a readable realization of an auditable temporal substrate.

8 Broader Implications

Beyond demonstrating the feasibility of patient-journey construction from discharge summaries, our findings reveal practical implications for clinical NLP and usage in real-world contexts.

Reusable supervision. Clinical institutions often have abundant unstructured records but limited resources for expert manual annotation, secure computing resources, and access to large proprietary models (Balasubramanian et al., 2025; Ntinopoulos et al., 2025; Grothey et al., 2025). CliniciCIRCA offers a practical route from raw text to task-specific supervision: instruction tuning on only 1,000 silver-labeled examples improved every model on extraction and summarization, and four of five on temporal tagging. These results suggest that modest pipeline-generated datasets can support smaller, locally deployable models without requiring expert labeling of an entire clinical corpus or complex schema development, as has been the case in prior research (Section 7.2).

The necessity of human-in-the-loop oversight. Reliable evaluation is particularly important in clinical NLP because infrequent errors can remain consequential even when aggregate performance appears high (Jiang et al., 2026; Hager et al., 2024). Prior work in psychiatric NLP similarly finds that LLM outputs may resemble expert responses on surface-level properties such as tone while diverging on clinically consequential dimensions (Chandra et al., 2025; Wang et al., 2025b). Our automated

judge agrees strongly with mental health clinicians on the large majority of correct events, but is less reliable in identifying the nature of rare errors, particularly those involving pre-admission medications and date granularity. Producing a trustworthy reference set therefore requires exhaustive clinician-guided adjudication, since expert review remains necessary to resolve ambiguous temporal decisions and determine which errors are clinically meaningful (Bavaresco et al., 2025; Diekmann et al., 2025; Chen et al., 2024).

Supporting longitudinal chart review. Reviewing longitudinal records is time- and effort-intensive, especially for multimorbid patients whose histories could be distributed across multiple clinical encounters and narrative documents (Van Veen et al., 2024). Our proposed Stage 3 renders the event-level timeline as a readable patient journey while retaining its calendar anchors and temporal uncertainty. Such summaries could provide clinicians with a traceable overview before time-constrained follow-up visits, especially in contexts in which manually reconstructing a complex history from several discharge summaries may be impractical. That said, we underscore that they should complement rather than replace the source record, with individual claims remaining connected to inspectable event-level evidence.

9 Conclusion and Future Work

CliniciCIRCA shows that recovering a patient journey from clinical narrative is best treated not as a single generation problem, but as separable representational decisions: what happened, when it happened, and how it should be rendered for review. This decomposition makes the resulting chronology inspectable, exposes where temporal ambiguity remains unresolved, and enables each stage to be evaluated independently. Our evaluation further shows that automated judges are useful for locating problematic outputs, but not for replacing expert adjudication of rare temporal errors. At the same time, the reconstructed journeys provide effective supervision for smaller open-weight models, suggesting that structured intermediate representations can serve not only as outputs, but also as reusable training resources. Future work will extend CliniciCIRCA to multi-document patient records, richer temporal representations, and additional note types, to support traceable longitudinal chart review and memory-aware clinical systems.

Limitations

Although we present a framework with clinician-designed and verified definitions and steps, we recognize several areas for improvement.

Representational scope. Our date-centered schema supports chronological ordering but may impose artificial timing on atemporal facts or persistent states, such as sex, family history, and long-term substance use. Future work should compare alternatives, including atemporal labels, validity intervals, and separate state/event categories, and evaluate their effects on timeline fidelity, summarization, and borderline cases.

Summarization salience. Stage 3 follows a clinician-authored specification for retaining, combining, and omitting events in the narrative summary. Because the same two clinicians developed this specification and conducted the faithfulness and temporal-accuracy evaluation described in Section 5.4, the results may understate disagreement in salience or interpretation. Independent evaluation would strengthen this validation. The specification also reflects one view of clinical relevance, since priorities may differ across clinicians and care settings. The resulting summaries should therefore not be treated as a single canonical account of a patient’s journey. Future work should include more diverse evaluators and clinical settings, salience policies, and multiple reference summaries.

Psychiatric symptom representation in Stage 3. Our Stage 3 specification (Appendix H) excludes affect descriptions with vital signs and routine examination findings to reduce boilerplate. As a result, mood and affect trajectories – a core component of psychiatric symptom tracking, as illustrated by the depression case in Section 5.5 – are preserved in the auditable Output A timeline (e.g., via diagnosis and treatment events) but not always carried into the reader-facing Output B narrative in their own right. Revisiting this exclusion for the mental-health setting specifically, for example by retaining clinically significant affect changes as a distinct category rather than folding them under general exam findings, is a natural next step for future research.

Single-document scope. CliniciCIRCA reconstructs the course described within one discharge summary per admission rather than multiple notes from the full longitudinal record. This bounded setting isolates narrative reconstruction, but cannot reconcile duplicated, conflicting, or evolving infor-

mation across admissions. Prior patient-journey methods are not directly comparable because they use different inputs, prompts, and context assumptions. We therefore compare 11 candidate models under the same framework (Section 4), but do not include a single-pass end-to-end baseline. Future work should ablate the modular design using the same LLM and extend it to multi-document records through cross-note entity resolution, temporal alignment, and representation of revisions in clinical understanding.

Despite these limitations, our proposed framework CliniciCIRCA provides a clinician-validated representation for open-vocabulary temporal reconstruction from raw clinical narratives.

Ethical Considerations

We use the MIMIC-III Clinical Database (v1.4), accessed through PhysioNet under its Credentialed Health Data Use Agreement (v1.5.0) with the required CITI training completed for all personnel with direct access or analysis of the data. LLMs were used in compliance with PhysioNet’s policy on responsible use of MIMIC data with LLMs and online services. We accessed Claude models through Amazon Bedrock and Gemini models through Vertex AI because the data were neither externally retained nor used for model training. All human evaluation was performed exclusively by the authors, without the recruitment of external human subjects. As a secondary analysis of de-identified records, this work is not human-subjects research and required no additional IRB review.

Because source notes may unevenly document social history, trauma, or substance use across patients – for instance by diagnosis, demographic group, or documentation era – CliniciCIRCA’s extraction and summarization stages could reproduce such documentation asymmetries rather than correct them; auditing this is an important direction for future work. We also note that pipeline errors are not equally consequential: a mis-tagged or fabricated event involving suicidality, self-harm, or medication timing carries substantially more clinical risk than a mis-tagged laboratory value. Given this, we emphasize that Stage 3 outputs (Output B) are intended to complement, not replace, the source record, and should remain linked to their inspectable Stage 1-2 source events (Output A) in any deployment setting within the healthcare system, particularly for safety-relevant content.

Acknowledgments

Zhang, Ding, and De Choudhury were partly supported through funds from The Children’s Healthcare of Atlanta Pediatric Technology Center at Georgia Tech. We thank Jiawei Zhou, Rijul Magu, Shravika Mittal, Viet Cuong Nguyen, Pinxian Lu, Owen Xingjian Zhang, Zeyu Hua, and Zikang Leng for valuable technical ideation and suggestions throughout this work. We are grateful to Franklin Ye Ruan for discussions of clinical and medical knowledge that informed the design of this study. We thank Sonakshi Sisodiya for early-stage annotation, and Andrew Zhao for technical and server support. We also thank Kaike Ping for statistical guidance and support.

References

- Abdelrahman Abouelenin and 1 others. 2025. Phi-4-Mini Technical Report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Ghada Alfattni, Niels Peek, and Goran Nenadic. 2020. [Extraction of temporal relations from clinical free text: A systematic review of current approaches](#). *Journal of Biomedical Informatics*, 108:103488.
- Ali Amirahmadi, Farzaneh Etminani, Jonas Björk, Olle Melander, and Mattias Ohlsson. 2025. [Trajectory-Ordered Objectives for Self-Supervised Representation Learning of Temporal Healthcare Data Using Transformers: Model Development and Evaluation Study](#). *JMIR Medical Informatics*, 13:e68138.
- Anthropic. 2025. [Claude Haiku 4.5 System Card](#). System card.
- Anthropic. 2026. [Claude Sonnet 4.6 System Card](#). System card.
- Jeya Balaji Balasubramanian, Daniel Adams, Ioannis Roxanis, Amy Berrington De Gonzalez, Penny Coulson, Jonas S. Almeida, and Montserrat García-Closas. 2025. [Leveraging large language models for structured information extraction from pathology reports](#). *Journal of Pathology Informatics*, 19:100521.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. [With Little Power Comes Great Responsibility](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.
- Centers for Medicare & Medicaid Services and National Center for Health Statistics. 2011. [ICD-9-CM Official Guidelines for Coding and Reporting](#). U.S. Department of Health and Human Services. Effective October 1, 2011.
- Stéphane Champely. 2020. [pwr: Basic Functions for Power Analysis](#). R package version 1.3-0.
- Mohit Chandra, Siddharth Sriraman, Gaurav Verma, Harneet Singh Khanuja, Jose Suarez Campayo, Zihang Li, Michael L. Birnbaum, and Munmun De Choudhury. 2025. [Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11083–11113, Albuquerque, New Mexico. Association for Computational Linguistics.
- Rochana Chaturvedi. 2024. [Temporal Knowledge Graph Extraction and Modeling across Multiple Documents for Health Risk Prediction](#). In *Companion Proceedings of the ACM Web Conference 2024*, pages 1182–1185, Singapore Singapore. ACM.
- Rochana Chaturvedi, Peyman Baghersshahi, Sourav Medya, and Barbara Di Eugenio. 2025. [Temporal relation extraction in clinical texts: A span-based graph transformer approach](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25765–25788, Vienna, Austria. Association for Computational Linguistics.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the Judge? A Study on Judgement Bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.
- J. Cohen. 1988. *Statistical power analysis for behavioral sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NY.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context,](#)

- and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.
- Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Miranda Schnier, Kyle Burton, Cris Ebby, Jillian Gorski, Matthew Kalscheur, Samy Khalil, Marie Pisani, Tyler Rubeor, Peter Stetson, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 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.
- Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H. Shah. 2025. [TIMER: temporal instruction modeling and evaluation for longitudinal clinical records](#). *npj Digital Medicine*, 8(1):577.
- Anne de Hond, Tuur Leeuwenberg, Richard Bartels, Marieke van Buchem, Ilse Kant, Karel GM Moons, and Maarten van Smeden. 2024. [From text to treatment: the crucial role of validation for generative large language models in health care](#). *The Lancet Digital Health*, 6(7):e441–e443.
- Yella Diekmann, Chase Fensore, Rodrigo Carrillo-Larco, Eduard Castejon Rosales, Sakshi Shiromani, Rima Pai, Megha Shah, and Joyce Ho. 2025. [LLMs as Medical Safety Judges: Evaluating Alignment with Human Annotation in Patient-Facing QA](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 217–224, Viena, Austria. Association for Computational Linguistics.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- George L. Engel. 1977. [The Need for a New Medical Model: A Challenge for Biomedicine](#). *Science*, 196(4286):129–136.
- Gabriel Frattallone-Llado, Juyong Kim, Cheng Cheng, Diego Salazar, Smitha Edakalavan, and Jeremy C. Weiss. 2024. [Using Multimodal Data to Improve Precision of Inpatient Event Timelines](#). In De-Nian Yang, Xing Xie, Vincent S. Tseng, Jian Pei, Jen-Wei Huang, and Jerry Chun-Wei Lin, editors, *Advances in Knowledge Discovery and Data Mining*, volume 14648, pages 322–334. Springer Nature Singapore, Singapore. Series Title: Lecture Notes in Computer Science.
- Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. M. Churpek, and Majid Afshar. 2022. [Summarizing patients’ problems from hospital progress notes using pre-trained sequence-to-sequence models](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Gemma Team. 2026. [Gemma 4 Technical Report](#). *Preprint*, arXiv:2607.02770.
- Bastian Grothey, Jan Odenkirchen, Adnan Brkic, Birgid Schömig-Markiefka, Alexander Quaas, Reinhard Büttner, and Yuri Tolkach. 2025. [Comprehensive testing of large language models for extraction of structured data in pathology](#). *Communications Medicine*, 5(1):96.
- Yohan Bonescki Gumiel, Lucas Emanuel Silva E Oliveira, Vincent Claveau, Natalia Grabar, Emerson Cabrera Paraiso, Claudia Moro, and Deborah Ribeiro Carvalho. 2022. [Temporal Relation Extraction in Clinical Texts: A Systematic Review](#). *ACM Computing Surveys*, 54(7):1–36.
- Kilem Li Gwet. 2008. [Computing inter-rater reliability and its variance in the presence of high agreement](#). *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.
- Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. [Evaluation and mitigation of the limitations of large language models in clinical decision-making](#). *Nature Medicine*, 30(9):2613–2622.
- Zehua Jiang, Haichao Chen, Yilan Wu, Yiming Qin, Chenyang Pei, Dian Zeng, Bin Sheng, and Tien Yin Wong. 2026. [Beyond multiple-choice questions: Rethinking evaluation frameworks for large language models for clinical medicine](#). *Intelligent Medicine*, 6(2):109–115.
- Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. [Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries](#). In *Proceedings of the ACM Web Conference 2024*, WWW ’24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.
- Alistair Johnson, Tom Pollard, and Roger Mark. 2015. [MIMIC-III Clinical Database](#).
- Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3(1):160035.
- Karthikeyan K, Raghuveer Thirukovalluru, and David Carlson. 2025. [ClinStructor: AI-Powered Structuring of Unstructured Clinical Texts](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2822–2836, Mumbai, India. The Asian Federation of Natural Language

- Processing and The Association for Computational Linguistics.
- Dayeon Ki, Marine Carpuat, Paul McNamee, Daniel Khashabi, Eugene Yang, Dawn Lawrie, and Kevin Duh. 2025. Linguistic nepotism: Trading-off quality for language preference in multilingual rag. *arXiv preprint arXiv:2509.13930*.
- Michelle Kang Kim, Carol Roupheal, John McMichael, Nicole Welch, and Srinivasan Dasarathy. 2024. [Challenges in and Opportunities for Electronic Health Record-Based Data Analysis and Interpretation](#). *Gut and Liver*, 18(2):201–208.
- Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, James T Teo, and Richard J B Dobson. 2024. [Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study](#). *The Lancet Digital Health*, 6(4):e281–e290.
- Maya Kruse, Shiyue Hu, N. Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth M. Goldberg, and Yanjun Gao. 2025. [Large language models with temporal reasoning for longitudinal clinical summarization and prediction](#). *Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing*, 2025:20715 – 20735.
- Sayantan Kumar, Shahriar Noroozizadeh, Juyong Kim, and Jeremy C. Weiss. 2026. [Text Knows What, Tables Know When: Clinical Timeline Reconstruction via Retrieval-Augmented Multimodal Alignment](#). *arXiv preprint*. Version Number: 1.
- Chanseo Lee, Kimon A. Vogt, and Sonu Kumar. 2024. [Prospects for AI clinical summarization to reduce the burden of patient chart review](#). *Frontiers in Digital Health*, 6:1475092.
- Artuur Leeuwenberg and Marie-Francine Moens. 2020. [Towards Extracting Absolute Event Timelines From English Clinical Reports](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2710–2719.
- Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. [HaluEval: A large-scale hallucination evaluation benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.
- Claudio D. G. Linhares, Daniel M. Lima, Jean R. Ponciano, Mauro M. Olivatto, Marco A. Gutierrez, Jorge Poco, Caetano Traina, and Agma J. M. Traina. 2023. [ClinicalPath: A Visualization Tool to Improve the Evaluation of Electronic Health Records in Clinical Decision-Making](#). *IEEE Transactions on Visualization and Computer Graphics*, 29(10):4031–4046.
- Tyler J Loftus, Jeremy A Balch, Jenna L Marquard, Jessica M Ray, Brian S Alper, Neeraj Ojha, Azra Bihorac, Genevieve Melton-Meaux, Gopal Khanna, and Christopher J Tignanelli. 2024. [Longitudinal clinical decision support for assessing decisions over time: State-of-the-art and future directions](#). *DIGITAL HEALTH*, 10:20552076241249925.
- Zhimeng Luo, Yuelu Ji, Abhibha Gupta, Zhuochun Li, Adam Frisch, and Daqing He. 2023. [Towards Accurate and Clinically Meaningful Summarization of Electronic Health Record Notes: A Guided Approach](#). In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5, Pittsburgh, PA, USA. IEEE.
- Nikita Makarov, Maria Bordukova, Papichaya Quengdaeng, Daniel Garger, Raul Rodriguez-Esteban, Fabian Schmich, and Michael P. Menden. 2025. [Large language models forecast patient health trajectories enabling digital twins](#). *npj Digital Medicine*, 8(1):588.
- Bertalan Meskó and Eric J. Topol. 2023. [The imperative for regulatory oversight of large language models \(or generative AI\) in healthcare](#). *npj Digital Medicine*, 6(1):120.
- Meta AI. 2024. [Llama 3.3 70B Instruct model card](#). Hugging Face model card. Accessed July 27, 2026.
- Mistral AI. 2025. [Mistral-Small-3.2-24B-Instruct-2506 model card](#). Hugging Face model card. Accessed July 27, 2026.
- Shravika Mittal, Erin Kasson, Layna Paraboschi, Eleanor Laufenberg, Jiawei Zhou, Patricia A. Cavazos-Rehg, Tanushree Mitra, and Munmun De Choudhury. 2026. [Follow the rules \(or not\): Community norms and AI-generated support in online health communities](#). In *Proceedings of the International AAAI Conference on Web and Social Media*. AAAI Press. To appear.
- Gandhimathi Moharasan and Tu-Bao Ho. 2019. [Extraction of Temporal Information from Clinical Narratives](#). *Journal of Healthcare Informatics Research*, 3(2):220–244.
- Shahriar Noroozizadeh, Sayantan Kumar, George H. Chen, and Jeremy C. Weiss. 2025. [Temporally annotated textual time series from PubMed Open Access clinical case reports](#).
- Vasileios Ntinopoulos, Hector Rodriguez Cetina Biefer, Igor Tudorache, Nestoras Papadopoulos, Dragan Odavic, Petar Risteski, Achim Haeussler, and Omer Dzemali. 2025. [Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation](#). *BMJ Health & Care Informatics*, 32(1):e101139.
- Amy L. Olex and Bridget T. McInnes. 2021. [Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be](#). *Journal of Biomedical Informatics*, 118:103784.

- OpenMeditron. 2025. [Meditron3-Qwen2.5-7B model card](#). Hugging Face model card. Accessed July 27, 2026.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qwen Team. 2026. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Tong Ruan, Liqi Lei, Yangming Zhou, Jie Zhai, Le Zhang, Ping He, and Ju Gao. 2019. [Representation learning for clinical time series prediction tasks in electronic health records](#). *BMC Medical Informatics and Decision Making*, 19(S8):259.
- Koustuv Saha, Dong Whi Yoo, Vedant Das Swain, and Munmun De Choudhury. 2026. [Life events as predictors of wellbeing outcomes](#). *npj Digital Public Health*, 1(1):5.
- Peter Sarvari and Zaid Al-fagih. 2025. [Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method](#). *JMIRx Med*, 6:e67661–e67661.
- Tom M Seinen, Jan A Kors, Erik M Van Mulligen, and Peter R Rijnbeek. 2025. [Using Structured Codes and Free-Text Notes to Measure Information Complementarity in Electronic Health Records: Feasibility and Validation Study](#). *Journal of Medical Internet Research*, 27:e66910.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. [Medgemma technical report](#). *Preprint*, arXiv:2507.05201.
- Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah M Preum. 2025. Regen: A reliable evaluation framework for generative event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12146–12168.
- Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. 2024. [EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339, Miami, Florida, USA. Association for Computational Linguistics.
- William F. Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. [Temporal Annotation in the Clinical Domain](#). *Transactions of the Association for Computational Linguistics*, 2:143–154.
- 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.
- Renato Umeton, Anne Kwok, Rahul Maurya, Domenic Leco, Naomi Lenane, Jennifer Willcox, Gregory A. Abel, Mary Tolikas, and Jason M. Johnson. 2024. [GPT-4 in a Cancer Center — Institute-Wide Deployment Challenges and Lessons Learned](#). *NEJM AI*, 1(4).
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. [2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text](#). *Journal of the American Medical Informatics Association*, 18(5):552–556.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature Medicine*, 30(4):1134–1142.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.
- Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. [Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation](#). *Nature Scientific Data*.
- Jing Wang, Xing Niu, Juyong Kim, Jie Shen, Tong Zhang, and Jeremy Weiss. 2025a. [MIMIC-IV-Ext-22MCTS: A 22 Millions-Event Temporal Clinical Time-Series Dataset with Relative Timestamp for Risk Prediction](#).
- Jing Wang and Jeremy C. Weiss. 2025. [A Large-Language Model Framework for Relative Timeline Extraction from PubMed Case Reports](#). *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2025:598–606.
- Liwei Wang, Qiuhao Lu, Rui Li, Sunyang Fu, and Hongfang Liu. 2024. [Wonder at Chemotimelines 2024: MedTimeline: An End-to-End NLP System for Timeline Extraction from Clinical Narratives](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 483–487, Mexico City, Mexico. Association for Computational Linguistics.

- Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. 2025b. [Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19999–20017, Suzhou, China. Association for Computational Linguistics.
- World Health Organization. 2025a. Mental disorders. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>. Accessed: 2026-07-13.
- World Health Organization. 2025b. Mental health. <https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response>. Accessed: 2026-07-14.
- Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, Jingxian He, Wenye Hua, Mingyu Jin, Guang Chen, Yang Zhou, Zhao Li, Trisha Gupte, Ming-Li Chen, Zahra Azizi, Qi Dou, Bryan P. Yan, and 7 others. 2025. [Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education](#). *Communications Medicine*, 6(1):27.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.

A Dataset Extraction and Profile

A.1 Mental Health Admissions Subset

We operationalize mental-health-coded admissions using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM): an admission qualifies if it contains at least one diagnosis code from Chapter 5, Mental Disorders (290–319) (Centers for Medicare & Medicaid Services and National Center for Health Statistics, 2011) (Centers for Medicare & Medicaid Services and National Center for Health Statistics, 2011). Table A1 breaks down the key categories.

Our framework uses only free-text discharge summaries from the MIMIC-III NOTEEVENTS table, which describe each patient’s hospital stay. We link each summary to structured MIMIC-III fields only to define the study cohort and provide date anchors: diagnosis codes from DIAGNOSES_ICD, date of birth from PATIENTS, and admission and discharge dates from ADMISSIONS. These structured fields are not used as model inputs or training labels. The discharge summaries are provided to the framework in their original form, without chunking, section detection, or text normalization.

Code range	Diagnostic category
290–294	Organic psychotic conditions
295–299	Other psychoses
300–316	Neurotic disorders, personality disorders, and other nonpsychotic mental disorders
317–319	Intellectual disabilities

Table A1: Subcategories within ICD-9-CM Chapter 5, Mental Disorders (codes 290–319) (Centers for Medicare & Medicaid Services and National Center for Health Statistics, 2011). An admission is included in the mental-health-coded cohort if it contains at least one ICD-9-CM diagnosis code within the Mental Disorders chapter (290–319)

A.2 Benchmark Input Text Profile

Because both model evaluation and human validation require comparison against the complete source discharge summary, input length determines the amount of narrative context that must be processed and reviewed. Table A2 summarizes the raw-text lengths of the 52-document benchmark. The summaries contain 1,774 words on average (median: 1,624; SD: 821), ranging from 437 to 3,583 words and totaling 92,222 words. Character counts show similar variation, with a mean of 11,871 characters and a range of 2,952 to 24,808.

Metric	Mean	Median	SD	Min	Max	Total
Characters	11,871	11,042	5,481	2,952	24,808	617,267
Words	1,774	1,624	821	437	3,583	92,222

Table A2: Raw-text profile of the 52 discharge summaries in the benchmark.

These statistics reflect the original discharge summaries supplied to the framework without chunking, section parsing, or text normalization, and illustrate the long-context processing and manual-review burden of the benchmark.

A.3 Token Usage for Silver-Standard Generation

To quantify the inference required to construct the silver-standard supervision used in Section 6, we recorded the API token-usage metadata returned by Gemini 2.5 Pro for the 1,000 admissions processed through the frozen three-stage framework. Across Stages 1 to 3, generation processed 45.13M input tokens and produced 16.70M output tokens, for a combined total of 61.84M tokens, or 61,836 tokens per admission on average. Stage 1 accounted for 25.66M tokens (41.5% of the total), Stage 2 for 19.52M (31.6%), and Stage 3 for 16.66M (26.9%). Table A3 provides the complete stage-level breakdown.

These counts quantify the inference used to generate the 1,000-example silver corpus in Section 6 and illustrate the computational demands of processing complete discharge summaries and their intermediate representations through a multi-stage long-context pipeline. The resulting corpus, however, can be reused across three downstream tasks and multiple open-weight student models, amortizing the initial teacher-generation cost. Of the reported input tokens, 16,177,254 were served from cached context; because cached tokens are a subset of input tokens, they are not added again to the total.

B Agent Candidate Selection

Following the selection of agents in Section 4.2, a candidate is eligible if it supports a native context window of at least 32K tokens which is sufficient to fit our P99 input without truncation. All candidates are evaluated with identical prompts and greedy decoding ($t = 0$). An open-weight candidate additionally has to be (i) downloadable under a license permitting research and on-premise inference and (ii) loadable on our deployment stack across four

Stage	Input	Output	Total
Stage 1: Atomic Clinical Event Extraction	20,464,371	5,198,803	25,663,174
Stage 2: Date & Certainty Tagging	9,794,222	9,723,468	19,517,690
Stage 3: Temporally Guided Summarization	14,874,581	1,780,523	16,655,104
Total	45,133,174	16,702,794	61,835,968

Table A3: Token usage for generating the 1,000-example silver-standard corpus with the frozen Gemini 2.5 Pro. Total tokens are the sum of input and output tokens. The grand total corresponds to an average of 61,836 tokens per admission across all three stages. Of the input tokens, 16,177,254 were served from cached context and are already included in the input counts.

L40S GPUs at bfloat16 precision. A proprietary candidate additionally has to be (iii) served through a managed endpoint compatible with our institutional data-use agreement (e.g., Amazon Bedrock or Google Vertex AI) and (iv) cost-feasible at corpus scale. A complete table of agent specifications can be found in Table A4.

Prior work recommends low decoding temperatures for classification and labelling tasks to stabilize outputs (Jin et al., 2024). Because our stages demand both high precision and high recall, including sharp inclusion/exclusion at extraction and unambiguous tagging decisions downstream, we set temperature $t = 0$ (greedy decoding) for all three agentic stages. All other generation settings were held constant across models, and for open-weight models this produces byte-identical outputs across runs.

C Benchmark Size Selection Via Power Analysis

We fix the number of manually evaluated documents using an a priori power analysis for a chi-square test (Cohen, 1988). The required sample size is

$$N = \frac{\lambda^*}{w^2}, \quad (1)$$

where w is Cohen’s effect-size index and λ^* is the noncentrality parameter at which the noncentral chi-square distribution $\chi^2(df, \lambda)$ reaches the target power relative to the critical value $\chi^2_{1-\alpha}(df)$.

For a pairwise comparison with one degree of freedom, we targeted a large effect size ($w = 0.5$), a significance level of $\alpha = 0.05$, and power of $1 - \beta = 0.95$. This gives

$$\lambda^* = \left(z_{1-\alpha/2} + z_{1-\beta} \right)^2 = 12.99, \quad (2)$$

and therefore

$$N = \frac{12.99}{0.5^2} = \frac{12.99}{0.25} = 51.98. \quad (3)$$

We round this value up to $N = 52$ documents. Applying a finite-population correction for the full corpus of 14,882 documents produced an adjusted estimate of $N_{\text{adj}} = 51.80$, leaving the required sample size effectively unchanged. At $N = 52$, the achieved power is 0.950.

The analysis is conducted using the pwr package (Champely, 2020) and independently verified using the noncentral chi-square distribution implemented in SciPy (Virtanen et al., 2020). The sampling unit is the discharge-summary document. Thus, the independence assumption applies across sampled documents, and correlations among events within a document do not artificially increase the effective sample size.

D Stage 1 & Stage 2 Generation Prompts

We present the concrete definitions for each of the sub-dimensions of timeline-robustness.

We provide the complete prompts used for Stages 1 and 2 in Tables A5 and A6. Each stage uses a fixed system prompt and a document-specific user-prompt template. The system prompts were applied unchanged across all candidate models. At inference time, only the bracketed placeholders were replaced with the corresponding discharge summary, extracted events, and admission anchors. No model-specific prompt tuning or manual pre-processing was performed.

D.1 Stage 1: Atomic Clinical Event Extraction

The Stage 1 prompt instructs the model to extract all clinical information from the discharge summary as atomic, self-contained events. The prompt prioritizes coverage and source faithfulness: repeated mentions are retained, negation is preserved, and the model is instructed not to correct the note,

Agent Model	Provider	Params (total/active)	Arch.
Gemini 2.5 Pro (Comanici et al., 2025)	Google (via Vertex AI)	Undisclosed	n/r
Gemini 2.5 Flash	Google (via Vertex AI)	Undisclosed	n/r
Claude Sonnet 4.6 (Anthropic, 2026)	Anthropic (via Bedrock)	Undisclosed	n/r
Claude Haiku 4.5 (Anthropic, 2025)	Anthropic (via Bedrock)	Undisclosed	n/r
Llama 3.3 70B (Meta AI, 2024)	Meta	70B	Dense
Qwen 3.6 35B-A3B (Qwen Team, 2026)	Alibaba	35B / 3B	MoE
Gemma 4 26B-A4B (Gemma Team, 2026)	Google	26B / 4B	MoE
MedGemma 27B (Sjellergren et al., 2026)	Google	27B	Dense
Mistral Small 3.2 24B (Mistral AI, 2025)	Mistral AI	24B	Dense
Meditron 3 (Qwen 2.5 7B) (OpenMeditron, 2025)	OpenMeditron/EPFL	7B	Dense
Phi-4 Mini 3.8B (Abouelenin et al., 2025)s	Microsoft	3.B	Dense

Table A4: Model agents included in the candidate pool. Models vary in provider, domain specialization, serving environment, parameter count, and architecture.

infer unstated information, or add clinical interpretation. Events are returned in their original document order as an unnumbered list. The complete prompt is shown in Table A5.

D.2 Stage 2: ISO Date and Certainty Tagging

The Stage 2 prompt assigns one temporal label to each event extracted in Stage 1. The model receives the full discharge summary as global context, together with the admission date, discharge date, date of birth, patient age, and the ordered list of atomic events. It is instructed to preserve the event text and assign exactly one of four labels: [EXACT], [APPROX], [PRE ADM], or [INDETERMINATE]. The prompt follows a conservative decision rule: explicitly written dates receive [EXACT], dates resolved from anchors or relative expressions receive [APPROX], events framed as prior history receive [PRE ADM], and [INDETERMINATE] is used only when no defensible temporal resolution is available. The model is explicitly prohibited from guessing dates. The complete prompt is shown in Table A6.

Type	Prompt
System Prompt	You are an atomic clinical event extraction engine. Your task is to extract a list of atomic clinical events (Presented as ATOMIC_CLINICAL_EVENTS) from an unstructured discharge summary. CLINICAL EVENTS: A clinical event is one clinically meaningful, patient-related unit of information documented in the note, expressed as a self-contained statement describing an action, condition, observation, or change in the patient's clinical course. An event may describe something that: 1. Happened 2. Was observed 3. Was reported 4. Was performed 5. Was denied 6. Was known about the patient This includes, but is not limited to: Diagnoses, symptoms and signs, Procedures, lab results, imaging findings, medications, social history, behavioral events, family history, clinical decisions and plans, negated findings (these must be preserved exactly). ATOMIC CLINICAL EVENTS: We define an atomic clinical event as the smallest unit of clinical information expressible as a single subject-verb-object clause that cannot be further decomposed without loss of clinical meaning. Instructions: 1. Expand all abbreviations but keep the original in parentheses (e.g., "shortness of breath (SOB)", "heart rate (HR) of 96"). 2. Extract every mention of every fact, even if the same fact appears in multiple sections. Do not deduplicate. 3. Do not correct errors, infer unstated information, or add clinical interpretation. Extract only what is explicitly written. 4. Negated findings must be preserved exactly (e.g., "The patient has no thyromegaly"). 5. Preserve the tense from the original text. For structured data without verbs (such as labs and vitals), use "was." 6. Output the atomic clinical events in document order as a list, with each item starting with "- ". Do not number the bullets. Do not add explanations. Do not add any extra text.
User Prompt	DISCHARGE_SUMMARY: <discharge_summary>

Table A5: Prompt used for atomic clinical event extraction

Type	Prompt
System Prompt	You are a temporal tagger for atomic clinical events. Assign exactly one temporal tag to each event in the given ATOMIC_CLINICAL_EVENTS list, producing the TAGGED_ATOMIC_CLINICAL_EVENTS list. Accurate tags let us reconstruct a patient’s timeline from an unstructured note, so a wrong date is worse than a coarse one — NEVER invent or guess a date. You are given the original note (DISCHARGE_SUMMARY) as global context and ANCHOR_DATES (admission date, discharge date, and patient age if stated). DATE FORMATS: - When a date is assigned, use the most specific format supported by the source: [YYYY-MM-DD], [YYYY-MM], or [YYYY]. - For events spanning a range, use [START to END], where START and END can each be in any of the above formats (e.g., [2000-01-01 to 2000-01-02], [1998-11 to 2001-08-02]). - Never add precision the source lacks: month-only resolves to [YYYY-MM]; year-only resolves to [YYYY]. TAGS: assign the one option that applies. [EXACT] – a date (or range) is written in the event text itself and applies to it. Use it as written, normalized, regardless of where it falls: a dated pre-admission event (e.g. a past surgery) is EXACT, not [PRE ADM]; a future follow-up date is EXACT. Examples include: - “[2000-01-01] [EXACT] The patient underwent a surgical procedure on 2000-01-01” - “[2010-03-04] [EXACT] On 2010-03-04 the patient was brought to the emergency department (ED)” - “[1980-05-13] [EXACT] The patient is to follow up with his primary care provider on 1980-05-13” (a written date is EXACT even when it falls after discharge) [APPROX] – no explicit date is written in the event string, the event resolves to an anchor (admission, discharge, or DOB). Covers: the admission/presentation encounter; a relative offset from an anchor, at the offset’s precision; an in-stay event with no specific date (use the [admission to discharge] range); a plan or instruction issued at discharge. Examples (admission 2010-01-06, discharge 2010-01-10): Examples include: - “[2010-01-01] [APPROX] 5 days before admission, the patient attempted suicide” - “[2003] [APPROX] The patient quit smoking 7 years ago” - “[2010-01-06 to 2010-01-10] [APPROX] A procedure was held during the hospitalization” - “[2010-01-06] [APPROX] In the emergency department (ED), the patient did well” (a prior related event established the ED visit was on 2010-01-06; events tied to that visit resolve to it) [PRE ADM] – the event is framed as prior to or continuous through this admission — including but not necessarily limited to past medical history, “history of . . .”, “status post”/“s/p”, prior events, family or social history with no exact or relative time mentions, demographics, or an undated chronic state. Carries no [DATE] tag in front. Examples include: - “[PRE ADM] The patient has a history of depression” - “[PRE ADM] The patient is a 37 year old female” - “[PRE ADM] The patient was involved in an accident” - “[PRE ADM] The patient’s mother died at age 40 of cancer”. Strict Guideline Prefer the three tags above. Use the last option, [INDETERMINATE], only as a last resort. [INDETERMINATE] – use ONLY after [EXACT], [APPROX], and [PRE ADM] have each been ruled out: no written date (not EXACT), no resolution to any anchor (not APPROX), and no indication it occurred before admission (not PRE ADM). The timing genuinely cannot be determined. This is not a fallback for uncertainty — if any of the first three plausibly applies, use that one. Carries no date. An example includes: - “[INDETERMINATE] An outpatient ultrasound was recommended at some point” Output Instructions 1. Preserve the exact wording of each input event in the output. Do not reword, reorder, correct errors, infer unstated facts, or add interpretation. 2. One bullet per event, starting with “- ”. Tag block is “[DATE] [TAG]” for EXACT and APPROX, and the tag alone (“[PRE ADM]” or “[INDETERMINATE]”) for the two that carry no date. 3. Do not number bullets. Do not add any text outside the bullets.
User Prompt	ANCHOR_DATES: - Date of Birth: date_of_birth - Admission Date: <admission_date> - Discharge Date: <discharge_date> DISCHARGE_SUMMARY: <discharge_summary> ATOMIC_CLINICAL_EVENTS: <atomic_clinical_events>

Table A6: Prompt used for date & certainty tagging

E Stage 1 & Stage 2 Evaluation Rubrics

Stage 2 is evaluated using four complementary dimensions. Completeness measures whether the extracted timeline captures the source-grounded clinical events in the discharge summary. Semantic faithfulness assesses whether each event preserves the meaning of the source, including negation, uncertainty, quantities, and other qualifiers. Temporal-tag accuracy evaluates both the assigned certainty category and the resolved date or date range. Anti-hallucination identifies events that introduce claims unsupported by the source note. It's important to note that anti-hallucination was evaluated only by human annotators. Unlike the other dimensions, this criterion requires determining whether any claim in an extracted event lacks support anywhere in the source note. Prior work shows that LLM-as-judge reliability varies substantially across tasks and evaluation properties, and that even strong LLMs remain imperfect at recognizing hallucinated content (Bavaresco et al., 2025; Li et al., 2023). This limitation is especially consequential in the clinical domain, where subtle, plausible unsupported claims may be difficult to distinguish from source-grounded information. Because a missed hallucination would directly contaminate the clinician-verified reference set, we reserved this criterion for exhaustive human review rather than treating an LLM's factuality judgment as ground truth.

All rubrics are included in Tables A7 to A10.

F Stage 1 & Stage 2 Evaluation Platform

Human evaluation of Stage-2 outputs was conducted through a purpose-built local web application (Figure A1). The tool runs as a self-contained Python program with a static HTML interface, requires no external services or database, and is accessed through SSH port forwarding so protected health information remains on the host machine. Model outputs are loaded into an ordered annotation queue, and annotators resume from their first unfinished document.

The interface displays the full discharge summary alongside the model-generated events in source order, including their temporal and certainty tags. For each event, annotators independently assess temporal accuracy, certainty-tag accuracy, semantic faithfulness, and hallucination, with criterion-specific comment fields for flagged errors. To capture omissions, the interface also records the

number of missing events and the corresponding source passages.

The platform includes progress tracking, bulk confirmation for all-correct events, checks that prevent advancing with incomplete ratings, keyboard navigation, and immediate crash-safe saving to per-annotator JSON files. Completed work can therefore be interrupted and resumed without loss.

G Stage 1 & Stage 2 Evaluation Agreement and Error Analysis

Table A11 reports agreement between the Gemini 3.1 Pro evaluator and clinician-guided human annotations across 15,891 Stage-2 events from 52 discharge summaries. We report Cohen's κ , Gwet's AC1, and positive and negative agreement. Because confirmed errors are rare relative to correct events, Cohen's κ is sensitive to class prevalence; Gwet's AC1 and the class-specific agreement measures provide complementary views of agreement under this imbalance.

Overall agreement is high on the dominant non-error class: negative agreement reaches 0.986, while Gwet's AC1 is 0.972. Agreement is lower on the rare error class, with positive agreement of 0.669. This indicates that the evaluator and human reviewers agree strongly on which events are correct, but differ more often when identifying or characterizing errors.

The disagreements are concentrated in temporal accuracy rather than semantic faithfulness. Although human annotation and the automated judge report nearly identical total numbers of temporal errors (608 versus 602), an event-level breakdown reveals a strong directional asymmetry: human review identified 175 over-specified tags on pre-admission medications and demographics, 169 of which the judge failed to flag, whereas judge-only findings were concentrated in overlapping date-range and date-granularity codes (Appendix G.1). Temporal-accuracy agreement reaches $\kappa = 0.617$, whereas semantic-faithfulness negative agreement is 1.000. Manual review identified two recurring sources of temporal disagreement. First, some models anchor home medications to the admission date using [APPROX], while clinicians classify them as [PRE ADM]; the automated evaluator does not consistently penalize this distinction. Second, the evaluator sometimes accepts a coarse date range even when the event can be resolved to a more specific

Component	Dimension A: Completeness
Guiding Question	Did the extraction capture the source-grounded events from the note?
Definition	A source-grounded clinical event is a single, self-contained assertion documented in the note and concerning the patient’s care. This includes, but is not limited to, conditions, findings, procedures, medications, laboratory and vital-sign values, plans and instructions, demographics, allergies, admission and discharge facts, and family or social history. An assertion is source-grounded when it is stated in the note; it does not need to be clinically salient. Clinical salience is not assessed under this rubric.
Exclusions	An assertion is excluded only when its entire content is administrative, such as a clinician or provider name alone, telephone or fax number, pager number, address, ward or room identifier, or document-formatting artifact. An administrative detail appearing within an otherwise clinical assertion does not exclude that assertion.
Reporting	Report every missing event, including its source span in the note and a one-line description. The expected-event set is the union of the source-grounded candidate events and the missing events reported by the evaluator. Coverage is calculated as the number of captured expected events divided by the size of this union.
Score 5	The candidate list captures at least 95% of the expected source-grounded clinical events.
Score 4	The candidate list captures at least 80% but less than 95% of the expected source-grounded clinical events.
Score 3	The candidate list captures at least 60% but less than 80% of the expected source-grounded clinical events.
Score 2	The candidate list captures at least 40% but less than 60% of the expected source-grounded clinical events.
Score 1	The candidate list captures less than 40% of the expected source-grounded clinical events.

Table A7: Stage-2 evaluation rubric for Dimension A: Completeness.

in-stay anchor. Thus, disagreement is concentrated in borderline cases of pre-admission status and date granularity rather than being distributed uniformly across the task.

The integrity of the gold set does not depend on evaluator agreement. Human annotators exhaustively reviewed all 15,891 events under clinician guidance and corrected every confirmed error after escalating errors to discussion with clinicians, regardless of whether the automated evaluator detected it. This process produced 52 human-corrected pairs of discharge summaries and tagged timelines.

G.1 Temporal Error-Type Analysis

To characterize the lower positive agreement on temporal accuracy, we compared the finalized event-level error codes assigned through exhaustive human review with those assigned by the calibrated LLM judge across the 52-note benchmark.

As shown in Figure A2, the aggregate counts are similar: human annotation identified 608 temporal errors, compared with 602 identified by the judge. Human review found at least one temporal error in 50 of the 52 notes, whereas the judge identified an error in 46.

The clearest asymmetry concerns over-specified temporal tags. Human annotation identified 175

such errors involving pre-admission medications (151 events) and patient demographics (24 events), of which the judge failed to flag 169 (96.6%). These are not cases in which the human and judge applied different error codes to the same behavior: the judge returned no temporal error at all. The pattern therefore represents a systematic false-negative class in which the judge accepts admission-anchored [EXACT] or [APPROX] labels for standing or pre-admission facts that should receive [PRE ADM].

The reverse asymmetry is qualitatively different. Of the 200 judge-only errors, 80 were coded as `date_range_wrong` and 56 as `date_too_coarse`. These codes describe closely related aspects of temporal granularity and may place the boundary between an incorrect range and an insufficiently precise date differently. Thus, similar aggregate error counts do not indicate equivalent error detection: human review exposes a clinically coherent error class that the judge largely misses, whereas much of the judge’s excess is concentrated around rubric boundaries for range and granularity.

G.2 Other Error-Types and Analysis

The 21 non-temporal errors are minor and mechanical, encompassing just 0.13% of all 15,891 events.

Component	Dimension B: Semantic Faithfulness
Guiding Question	Does each candidate event preserve the clinical meaning of the source note while allowing the permitted normalizations?
Scope	Judge only the event text. Ignore the bracketed date and temporal-tag block preceding it, because temporal tags are evaluated separately. Within this dimension, assess only whether temporal expressions written inside the event text match those written in the source note.
Definition	An event preserves the source meaning only when all of the following properties survive rewriting: polarity or negation; epistemic status, including hedging and expressions such as “likely” or “concern for”; experiencer, such as the patient versus a family member; quantities, units, doses, and laterality; severity, chronicity, and frequency; temporal expressions appearing in the text; and conditionality or hypotheticality. Each event receives a binary semantic judgment. A wording change that preserves all meaning-invariant properties is not an error.
Hallucinations	Events with no identifiable source assertion are hallucinations and are evaluated under Dimension D. They are excluded from the denominator for Dimension B. No event may be counted as an error under both Dimensions B and D.
Reporting	Report every event that fails, including its event ID, corresponding source span, and one error code from the following list: • negation_lost: a negated finding is asserted, or an asserted finding is negated. • uncertainty_lost: a hedged claim is asserted as fact, or a factual claim is incorrectly hedged. • experiencer_shifted: a statement concerning a family member or another person is attributed to the patient, or vice versa. • value_or_unit_altered: a quantity, unit, dose, count, or laterality differs from the source. • qualifier_altered: severity, chronicity, or frequency differs from the source. • temporal_cue_altered: a temporal expression in the event text differs from or is omitted relative to the source. • conditionality_lost: a conditional or hypothetical statement is presented as an actual occurrence. • abbreviation_misexpanded: an abbreviation is expanded to a meaning unsupported by the source. • other_distortion: the meaning differs for a reason not covered by the other error codes.
Normalization	Separately record events whose wording deviates from the permitted normalizations while preserving all meaning-invariant properties using the flag normalization_deviation. These cases are not semantic errors and do not affect the score.
Score 5	At least 95% of candidate events preserve the source meaning.
Score 4	At least 80% but less than 95% of candidate events preserve the source meaning.
Score 3	At least 60% but less than 80% of candidate events preserve the source meaning.
Score 2	At least 40% but less than 60% of candidate events preserve the source meaning.
Score 1	Less than 40% of candidate events preserve the source meaning.

Table A8: Stage-2 evaluation rubric for Dimension B: Semantic Faithfulness.

Component	Dimension C: Temporal-Tag Accuracy
Guiding Question	For each candidate event, is the assigned temporal tag correct, and is the assigned date correct?
Tag Format	Every event carries exactly one temporal tag. [EXACT] and [APPROX] also carry a date or date range. [PRE ADM] and [INDETERMINATE] carry no date.
[EXACT]	A date is written in the event text itself. A date whose month and day are written and whose year is resolved from the note is also [EXACT]. For example, “on 5-9” in a note with admission year 2135 becomes [2135-05-09] [EXACT]. An event whose text contains no date is never [EXACT].
[APPROX]	No date is written in the event text, but the event resolves to the admission date, discharge date, date of birth, a relative offset from one of these anchors, or a date established elsewhere in the note. An in-stay event without a specific date receives the admission-to-discharge range. A plan issued at discharge receives the discharge date.
[PRE ADM]	The event is framed as occurring before or continuing through the current admission, with no written or resolvable date. This includes history, “status post” statements, demographics, undated chronic states, and undated family or social history.
[INDETERMINATE]	None of the other categories applies and the event’s timing cannot be determined. This is a last-resort label.
Date Precision	A date is correct only when it matches the date that should have been assigned at the same level of precision. Assigning [2135-05] when [2135-05-09] was recoverable is an error, as is assigning [2135-05-09] when only [2135-05] was supported. For a date range, both endpoints must match.
Event Score 1.0	Assign 1.0 when the temporal tag is correct and, for [EXACT] and [APPROX], the date is also correct. [PRE ADM] and [INDETERMINATE] receive 1.0 when the tag is correct because they carry no date.
Event Score 0.5	Assign 0.5 only when both the assigned and correct tags are [EXACT] or [APPROX], and exactly one of the tag or date is correct. This includes cases where the tag is correct but the date is wrong, or the date is correct but [EXACT] and [APPROX] are interchanged.
Event Score 0.0	Assign 0.0 in every other case. This includes assigning a no-date tag when a date-bearing tag is correct; assigning a date-bearing tag when a no-date tag is correct; confusing [PRE ADM] with [INDETERMINATE]; assigning both the wrong tag and wrong date; or emitting [EXACT] or [APPROX] without a date.

Table A9: Stage-2 evaluation rubric for Dimension C: Temporal-Tag Accuracy (Part 1 of 2).

Component	Dimension C: Temporal-Tag Accuracy
Reporting	Report every event receiving a score below 1.0. Include its event ID, the temporal tag and date believed to be correct, and one error code from the following list: • tag_over_specific: [EXACT] or [APPROX] was assigned where [PRE ADM] or [INDETERMINATE] was correct, meaning that a date was invented. • tag_under_specific: [PRE ADM] or [INDETERMINATE] was assigned where a date was recoverable. • tag_exact_approx_confused: [EXACT] and [APPROX] were interchanged. • tag_preadm_indet_confused: [PRE ADM] and [INDETERMINATE] were interchanged. • date_wrong_value: the assigned date contains the wrong day, month, or year. • date_too_coarse: the date is correct but less precise than the note supports. • date_too_precise: the date is more precise than the note supports. • date_range_wrong: one or both endpoints of a date range are incorrect. • date_missing: [EXACT] or [APPROX] carries no date. • date_wrongly_present: [PRE ADM] or [INDETERMINATE] incorrectly carries a date. • other_temporal_inaccuracy: another temporal error, accompanied by an explanation.
Aggregate Rate	Temporal-tag accuracy is calculated as the sum of the per-event scores divided by the number of assessed events, expressed as a percentage.
Score 5	At least 95% temporal-tag accuracy.
Score 4	At least 80% but less than 95% temporal-tag accuracy.
Score 3	At least 60% but less than 80% temporal-tag accuracy.
Score 2	At least 40% but less than 60% temporal-tag accuracy.
Score 1	Less than 40% temporal-tag accuracy.
Rate Calculation	For Dimensions A–C, calculate rates to one decimal place. Do not round the rate before assigning its score band. Report the counts used in the calculation.

Table A9: Stage-2 evaluation rubric for Dimension C: Temporal-Tag Accuracy (Part 2 of 2).

Component	Dimension D: Anti-Hallucination
Guiding Question	Does every factual claim in the candidate events come from the source note?
Definition	This dimension detects information added by the extraction. It does not assess information that was changed from an existing source assertion, and it does not penalize permitted rewrites.
Decision Rule	For each candidate event, determine whether the source note contains an assertion of which the event is a version, even if the event distorts that assertion. If such an assertion exists, the event is not a hallucination and any meaning error belongs under Dimension B. If no corresponding source assertion exists, the event is a hallucination.
Hallucination Types	• fabricated_event: an action, finding, procedure, or occurrence that the source note never states. • fabricated_attribute: a value, unit, dose, count, side, or qualifier attached to a real event when the source gives no such attribute. When the source gives a different value for a known event, the error belongs under Dimension B. • fabricated_entity: a medication, test, diagnosis, or person that the source note never names. • fabricated_date: a date written within the event text that does not appear in the source note. Errors in the bracketed date or temporal-tag block belong under Dimension C. • unwarranted_inference: a causal, diagnostic, or prognostic conclusion not explicitly stated in the source note. Clinical reasoning stated in the source may be extracted, but unsupported reasoning introduced by the model may not.
Do Not Assess	Do not assess temporal tags or bracketed dates, which belong under Dimension C; meaning changes to events that have an identifiable source, which belong under Dimension B; or any permitted rewrite.
Reporting	Report every hallucinated event, including its event ID and complete event text.
Score 1	No candidate event contains a hallucination.
Score 0	At least one candidate event contains a hallucination.

Table A10: Stage-2 evaluation rubric for Dimension D: Anti-Hallucination.

Unlike the temporal errors, almost none reflect clinical misunderstanding. The dominant failure mode is over-precision: the model quietly hardens hedged or masked source text into confident assertions. Roughly half of the semantic errors convert uncertainty into fact (e.g., “? <clinical event>” becomes “possible <clinical event>”), or add unsupported specificity to medications (such as “twice a day”, “twice daily” appended to a drug the note never dosed). The rest stem from misreading MIMIC’s de-identification masks and OCR-garbled tokens. The 7 hallucinations are similar in spirit — mostly home/admission medications and lab values carrying detail not fully grounded in the source — and the single completeness miss (1 missing patient date of birth event across 52 notes) is negligible. In short, on everything except temporal tagging the model is essentially faithful; its only residual weakness is a tendency to resolve ambiguity in the source rather than preserve it.

H Stage 3 Generation Prompt

The Stage 3 prompt converts the tagged event timeline into a chronological prose summary. It specifies how temporal tags should be rendered, which clinical information should be included or ex-

Dimension	Cohen’s κ	AC1	Pos.	Neg.
Overall	0.655	0.972	0.669	0.986
Temporal accuracy	0.617	0.971	0.631	0.986
Semantic faithfulness	0.611	0.999	0.611	1.000

Table A11: Agreement between the Gemini 3.1 Pro and gold annotation, calculated over 52 discharge summaries producing 15,891 events. AC1 denotes Gwet’s AC1; Pos. and Neg. denote positive and negative agreement on the rare error and dominant non-error classes, respectively. Anti-hallucination is excluded because it was evaluated only by human annotators.

cluded, how duplicate or overlapping events should be combined, and how factual and temporal fidelity should be preserved. Because of its length, the complete prompt is presented across three consecutive sub-tables (Table A12).

I Stage 3 Evaluation Rubrics

Stage 3 is evaluated by both human annotators and the calibrated LLM evaluator using two shared dimensions. Faithfulness to Source measures whether each clinical claim in the generated summary is supported by the human-verified event timeline and preserves its original clinical meaning.

Type	Prompt
System Prompt	You will convert a tagged clinical event list into a temporally structured prose summary. INPUTS You are given two inputs: 1. ORIGINAL DISCHARGE SUMMARY — the source note. Use it only to clarify wording, expand abbreviations, or resolve unclear phrasing. Do not take any fact from it that is not present in the tagged event list. 2. TAGGED EVENT LIST — the authoritative source. Every fact and every date in the summary must come from this list. TEMPORAL RENDERING RULES (mandatory, no exceptions) Each event carries a temporal tag. Render each tag exactly as specified: • [YYYY-MM-DD] [EXACT]: “On YYYY-MM-DD, ...” • [YYYY-MM-DD] [APPROX]: “Approximately on YYYY-MM-DD, ...” • [YYYY-MM-DD to YYYY-MM-DD] [APPROX]: “Approximately from YYYY-MM-DD to YYYY-MM-DD, ...” • [YYYY] [EXACT]: “In YYYY, ...” • [PRE ADM]: “Before admission, ...” Rules: • Write all dates in ISO format (YYYY-MM-DD), zero-padded, exactly as given in the tag. • The word “approximately” applies to the date, not to the event. The event is certain; only its timing is approximate. Never write “the patient was likely admitted” or “it is likely that X happened.” • Reproduce each event’s negations and hedges exactly as written; do not add, delete, or replace them. • If the event text restates timing already supplied by the tag, such as “admitted on 2122-5-19” or “on postoperative day two,” delete the repeated timing phrase from the sentence. • You may adjust grammar and tense minimally so that the sentence reads correctly after the temporal prefix. Do not change the clinical content. DEDUPLICATION AND MERGING • Exact duplicates: If two events have the same tag and the same content, output the content once. • Subsumption: If one event’s content is fully contained in another’s, such as “The patient is female” and “The patient is a 19-year-old female,” output only the more specific event. • Merging: Events with identical tags (the same date and the same type) must be combined under one temporal prefix and joined into one or a few sentences. Never repeat the prefix, and never merge events across different dates or types. This includes recombining events split from one clinical statement, such as “Her incision was clean,” “was dry,” and “was intact.” • Merge date of birth, age, and sex into one opening demographic sentence.

Table A12: Prompt used for temporal-guided clinical summarization (Part 1 of 3).

Type	Prompt
System Prompt	INCLUDE Include only the following categories. If an event does not fall into one of these categories, omit it. 1. Patient demographics: date of birth, age, and sex. 2. The reason for admission: the clinical problem or finding that led to the admission, and any diagnostic workup or counseling directly related to it. 3. In-stay events, limited to the following types: a. Procedures and surgeries performed. b. Intraoperative findings and repairs. c. Admissions to and transfers between services or units. d. Intubation, extubation, and other airway or ventilation events. e. Placement or removal of devices, lines, tubes, or catheters. f. Treatments or interventions started or stopped. g. Diagnoses made or revised during the stay. h. Consultations obtained. 4. The discharge: date of discharge, disposition, discharge condition, and follow-up instructions. 5. Social history relevant to the patient's mental health, unless it is described as non-contributory or insignificant in the source material. 6. Medications on admission and medications at discharge must each be rendered as a line-separated list under their own temporal prefix, with one medication per line copied verbatim from the event text. These are the only lists permitted in the output; all other content must be continuous prose, as in the following example: Before admission, the patient was taking: • Medicine A at X dosage • Medicine B Approximately on [discharge date], the patient was discharged on: • Medicine C for Y duration • Medicine D via route Z EXCLUDE Exclude the following categories entirely: 1. Laboratory values and laboratory results. 2. Vital-sign values. 3. Physical-examination findings, including normal and benign findings and descriptions of body habitus or affect. 4. Statements that a parameter or status remained stable, unchanged, or within normal limits. 5. Past medical history, past surgical history, and prior events that are not the reason for the current admission, unless the condition actively complicated the hospital course, required a consultation, or significantly altered medical management during the stay. 6. Allergies. FIDELITY • Use only facts present in the tagged event list. • Do not infer, interpret, diagnose, or add clinical judgment. • Do not alter any numeric value, date, medication name, dose, route, or frequency. • Preserve negations exactly. "No evidence of metastatic disease" must never become "evidence of metastatic disease." • Retain de-identification placeholders, such as [Last Name (STitle) 9125], exactly as written. Do not expand, guess, or replace them. • Do not add a heading, preamble, commentary, or closing statement.

Table A12: Prompt used for temporal-guided clinical summarization (Part 2 of 3).

Type	Prompt
System Prompt	ORDER AND FORMAT • Output continuous prose, except for the two medication lists specified above. Do not use other bullet points, headings, or tables. • Open with the demographic sentence. • Order all remaining content chronologically by the start date of its tag. For events sharing a start date, preserve the order in which they appear in the tagged event list. • Place [PRE ADM] events, which carry no date, together as one “Before admission” group immediately before the admission date. • Start a new paragraph at each of the following points: before admission, the admission date, and the discharge date. WORKED CONVERSIONS These examples demonstrate temporal rendering only. Whether an event is retained is determined by the INCLUDE and EXCLUDE rules above. Input: - [2000-01-01] [EXACT] The patient was admitted on 2000-01-01. Output: On 2000-01-01, the patient was admitted. Input: - [2000-01-03] [APPROX] She was extubated on postoperative day three. Output: Approximately on 2000-01-03, she was extubated. Input: - [2000-05-10 to 2000-05-30] [APPROX] The patient was admitted to the ICU service. Output: Approximately from 2000-05-10 to 2000-05-30, the patient was admitted to the ICU service. Input: - [PRE ADM] The patient was found to be unconscious. Output: Before admission, the patient was found to be unconscious. Input: - [2011] [EXACT] The patient is status post knee operation in 2011. Output: In 2011, the patient underwent knee operation. Merging two events sharing one tag: Input: - [2013-12-31] [APPROX] The patient’s social service worker recommended therapy. - [2013-12-31] [APPROX] The patient’s social service worker discussed treatment plans with the patient’s PCP. Output: Approximately on 2013-12-31, the patient’s social service worker recommended therapy and discussed treatment plans with the patient’s PCP. Subsumption: Input: - [PRE ADM] The patient is female. - [PRE ADM] The patient is a 58-year-old female. Output: Keep only the second event and incorporate it into the opening demographic sentence. Now produce the summary.
User Prompt	ORIGINAL_DISCHARGE_SUMMARY: <discharge_summary> TAGGED_EVENT_LIST: <tagged_event_list>

Table A12: Prompt used for temporal-guided clinical summarization (Part 3 of 3).

Menu
subject
HADM: 1 / 52
MODEL Gemini_2_5_Pre
split: benchmark52
STAGE2 EVENTS 261
Save now
Next

events rated 261 / 261 · COMPLETE ✓
Mark all remaining as correct (Yes)

EHR Note (Discharge Summary)

Stage 2 Tagged Events
Time tag correct? · Type tag correct? · Semantically faithful? · Hallucinated?

#1
TIME Yes No TYPE Yes No SEMANTIC Yes No
HALLUCINATED? Yes No

#2
TIME Yes No TYPE Yes No SEMANTIC Yes No
HALLUCINATED? Yes No

#3
TIME Yes No TYPE Yes No SEMANTIC Yes No
HALLUCINATED? Yes No

#4
TIME Yes No TYPE Yes No SEMANTIC Yes No
HALLUCINATED? Yes No

How many missing events total?
1
Default 0 — set to how many events you think the model failed to extract.

Any missing event

Other comments

Figure A1: Web interface of the Stage-2 event-level human annotation platform. A single subject is shown: the source discharge summary (left) beside the model’s Stage-2 tagged event list (right), with per-event Yes/No controls for the four rubric dimensions (time tag, type tag, semantic faithfulness, hallucination) and a footer for logging missed events. All protected health information and note/event text are redacted (gray bars) for this figure.

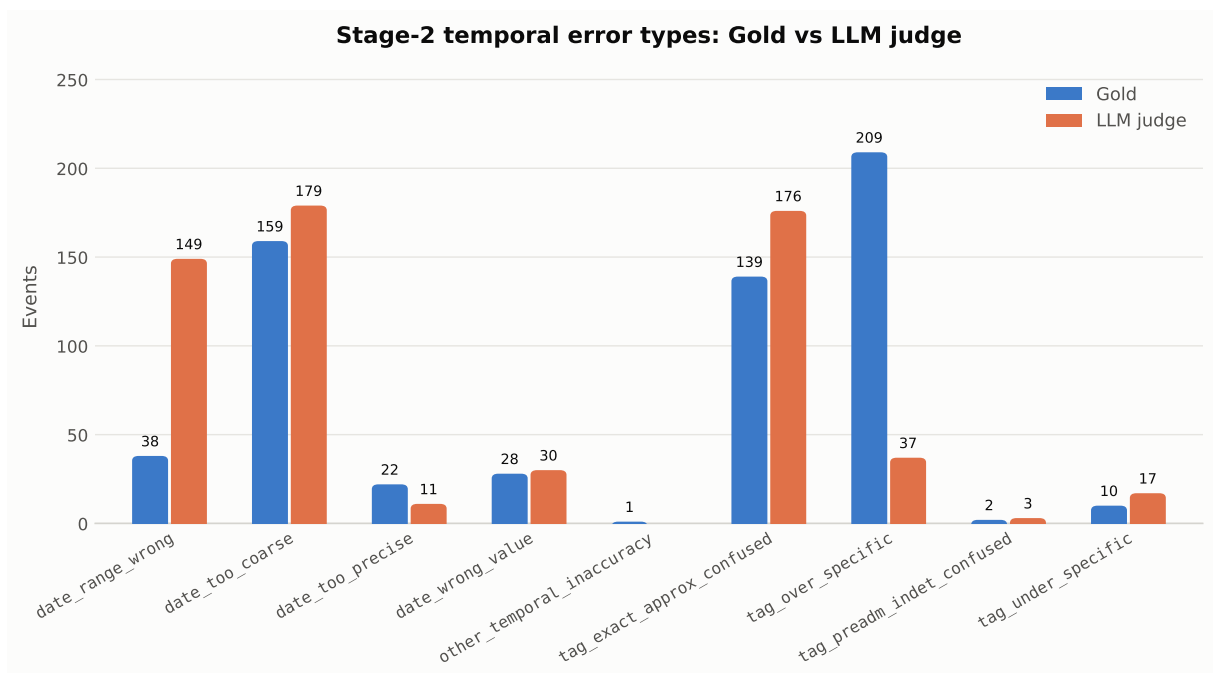


Figure A2: Stage-2 temporal-error counts from human annotation and the calibrated LLM judge. Despite similar totals (608 versus 602), human review surfaced 175 over-specified tags on pre-admission medications and demographics that the judge almost entirely missed, while judge-only errors concentrated in overlapping date-range and granularity categories.

Temporal Accuracy measures whether included events are rendered under the correct date and certainty tag and whether only events with identical

temporal tags are grouped together. The human-verified timeline serves as the reference of record; the original discharge summary is used only to

clarify unclear wording. Because selective content inclusion is expected in summarization, omitted events are not penalized under either dimension. The complete rubrics are provided in Tables A13 and A14.

J Stage 3 LLM and Auto-Evaluator Results

As shown in Table A15, faithfulness is saturated: all 52 summaries score band 5, with 5 distorted claims overall.

Temporal accuracy shows the mean likert band at 4.75 / 5, macro-mean 95.33% of events correctly placed (micro 96.17%, 257 errors overall; median 100%).

MEDCON precision (0.725) exceeds recall (0.519), F1 0.599. Asserted content is largely grounded, consistent with the faithfulness band; roughly half of source concepts are omitted. The judge scores only claims that were made and is therefore blind to this coverage gap, while MEDCON is insensitive to temporal placement. The two measure disjoint failure axes and are reported together for that reason.

K Stage 3 Evaluation Agreement

K.1 Annotation Protocol and Unit Definitions

Human annotators reviewed all 52 benchmark summaries alongside the judge's output. The review interface presents the judge's findings as a list and allows the annotator to mark any finding not a mistake.

K.2 Summary-level Detection Agreement

Table A17 treats each summary as one unit and asks whether the judge and the annotator agree that it contains at least one error. This framing avoids committing to a claim inventory as the denominator, and at a positive rate of 35% the prevalence index falls to 0.40, so Cohen's κ and Gwet's AC1 no longer diverge and both are interpretable. The judge flagged 18 summaries; the annotator's adjudication left at least one surviving flag in 13 of them and eliminated all flags in 5.

K.3 Error-count concordance

Table A18 compares the number of errors the judge reports in each summary against the number the annotator upheld. The two disagree sharply in magnitude and agree well in order. Lin's concordance correlation coefficient is 0.103 and ICC(A,1) is

0.105, both near the floor, because the judge over-reports by a mean of 4.3 errors per summary with 95% limits of agreement spanning -25.7 to $+34.3$ – an interval wider than the error count of any but the most heavily flagged summary. Spearman's ρ is 0.768, however: the judge ranks summaries by error burden much as the annotator does.

The gap between Spearman's ρ (0.768) and Pearson's r (0.511) reflects a single influential summary, subject 7676, at 100 judge errors against 5 upheld; rank-based statistics absorb it and moment-based ones do not. Taken together these results characterize the judge as a triage and ranking instrument rather than a measurement one. Its absolute error counts should not be reported as validated error rates, and any downstream comparison between generation models should use the judge's ordering rather than its magnitudes.

L Stage 3 Evaluation Platform

For Stage 3, we built a self-contained, locally-run web application for human review and correction of Gemini 2.5 Pro-generated Stage 3 summaries.

The review screen has three regions (Figure A3). The left panel shows the source material, toggling between the original discharge summary and the Stage 2 gold event list, and is read-only. The right panel presents the model's Stage 3 summary in an editable field: rather than only flagging problems, the reviewer corrects the summary in place to produce a gold summary. Edits autosave, and an "edited" indicator with a revert-to-original control distinguishes corrected from untouched summaries; the correction is stored as a separate `stage3_gold_summary` field, preserving the model's original output for comparison. The bottom panel surfaces an LLM-judge evaluation as a scaffold for verification. The summary is scored on two rubric criteria as listed in Appendix I.

M Qualitative Analysis Cases

In this section, we analyze three clinician-verified gold cases using WHO categories of mental disorders, symptoms, and risk factors as mentioned in Section 5.5, grounding each assignment in the gold timeline and treating unestablished links as co-documented rather than causal.

Component	Dimension A: Faithfulness to Source
Guiding Question	Is every clinical event and fact in the summary traceable to the source discharge summary or the human-verified event timeline, with no invented content, and is its clinical meaning preserved? Temporal placement is evaluated separately under Dimension B.
Reference	Evaluate the candidate summary primarily against the human-verified event timeline. The source discharge summary may be used only to clarify an event's wording when the timeline is unclear. Do not reward or penalize content that appears in the source note but is absent from the verified timeline.
Faithful Claim	A claim is faithful when it is traceable to a corresponding event and preserves the event's clinical meaning. Permitted rewriting, such as expanding an abbreviation while retaining the original or supplying a subject and verb for structured content, is not an error when the meaning remains unchanged. The summary must include all mental health-related and social history events from the original source.
Untraceable Claims	A claim is unfaithful when no supporting event exists in the human-verified timeline or source note. This includes an invented event, fact, medication, diagnosis, or clinical detail.
Meaning Distortions	A claim is also unfaithful when it corresponds to a real event but changes its clinical meaning. Examples include: • adding, dropping, or reversing a negation; • adding or removing a hedge or uncertainty marker; • changing a numeric value, dose, medication name, route, frequency, or laterality; • introducing an unsupported causal explanation, diagnosis, or clinical rationale; or • merging events into a new claim that neither source event supports.
Do Not Assess	Do not penalize omitted events, because selective inclusion is expected in a summary. Do not assess wrong dates, temporal tags, or temporal grouping under this dimension; these are evaluated under Dimension B.
Rate	The faithfulness rate is the percentage of clinical claims in the summary that are both traceable to the reference and meaning-preserving.
Score 5	At least 95% of the summary's clinical claims are faithful.
Score 4	At least 80% but less than 95% of the summary's clinical claims are faithful.
Score 3	At least 60% but less than 80% of the summary's clinical claims are faithful.
Score 2	At least 40% but less than 60% of the summary's clinical claims are faithful.
Score 1	Less than 40% of the summary's clinical claims are faithful.

Table A13: Stage-3 evaluation rubric for Dimension A: Faithfulness to Source.

Component	Dimension B: Temporal Accuracy
Guiding Question	Is every clinical claim in the summary placed under the same temporal tag assigned in the human-verified event timeline, and are only events with identical tags grouped together?
Scope	Evaluate temporal placement and grouping only. Whether an event is supported by the reference and whether its clinical meaning is preserved are assessed under Dimension A.
Temporal Tags	Every event carries one of the following temporal tags: • [DATE] [EXACT], where DATE is a single ISO date or ISO date range; • [DATE] [APPROX], where DATE is a single ISO date or ISO date range; or • [PRE ADM], for an event occurring before admission without a resolved date.
Valid Grouping	A grouping is valid only when all events within it share an identical temporal tag: the same date or date range and the same certainty marker. Events with the same date but different [EXACT] and [APPROX] markers must not be merged.
Misplacement	A claim is temporally misplaced when it appears under a different date, date range, or certainty tag from the one assigned in the verified timeline. Examples include placing a discharge-day event on the admission date, rendering an [EXACT] event as a date range, or assigning a date to a [PRE ADM] event.
Invalid Grouping	A temporal grouping is incorrect when: • events with different dates or date ranges are merged; • [EXACT] and [APPROX] events are merged despite sharing the same date; • a [PRE ADM] event is merged with a dated event; or • events sharing one identical tag are unnecessarily split across multiple temporal prefixes.
Do Not Assess	Do not assess invented or meaning-distorted events under this dimension; these are evaluated under Dimension A. Do not penalize omitted events. The order of correctly tagged groups is not an error, provided each group appears under its correct temporal prefix.
Rate	The temporal-accuracy rate is the percentage of included clinical events that are both correctly placed and correctly grouped. Rates are calculated to one decimal place, without rounding before assigning the score band. The counts used to calculate the rate must also be reported.
Score 5	At least 95% of included events are correctly placed and grouped.
Score 4	At least 80% but less than 95% of included events are correctly placed and grouped.
Score 3	At least 60% but less than 80% of included events are correctly placed and grouped.
Score 2	At least 40% but less than 60% of included events are correctly placed and grouped.
Score 1	Less than 40% of included events are correctly placed and grouped.

Table A14: Stage-3 evaluation rubric for Dimension B: Temporal Accuracy.

Model (summaries)	Faithfulness	Temporal Accuracy	MEDCON		
	Band / %	Band / %	P	R	F1
Gemini 2.5 Pro	5.00 / 99.93	4.75 / 95.33	0.725	0.519	0.599

Table A15: Evaluation results for summaries generated by Gemini 2.5 Pro. Faithfulness and temporal accuracy are reported as mean rubric bands and percentage scores. MEDCON precision, recall, and F1 measure clinical-concept overlap.

Metric	Model	SILVER→SILVER			SILVER→GOLD			GOLD		
		Zero	Few	IT	Zero	Few	IT	Zero	Few	IT
ROUGE-1	Gemma4-31B	.826±.008	.823±.008	.886±.006	.805±.011	.794±.013	.865±.013	–	.743±.026	.729±.030
	Qwen3.6-35B-A3B	.749±.010	.791±.009	.879±.007	.717±.014	.754±.016	.867±.011	–	.670±.036	.656±.037
	Qwen3-30B-A3B	.737±.013	.750±.013	.843±.012	.713±.014	.765±.012	.811±.018	–	.136±.040	.296±.041
	Mistral-24B	.736±.011	.761±.011	.878±.007	.679±.021	.724±.017	.864±.011	–	.698±.022	.763±.027
	Meditron-7B	.555±.021	.595±.022	.804±.017	.471±.030	.531±.031	.802±.021	–	.178±.043	.230±.031
ROUGE-2	Gemma4-31B	.711±.010	.693±.011	.804±.009	.675±.013	.649±.017	.776±.015	–	.607±.024	.629±.028
	Qwen3.6-35B-A3B	.621±.011	.671±.011	.797±.009	.575±.016	.622±.018	.782±.012	–	.555±.032	.551±.034
	Qwen3-30B-A3B	.598±.012	.617±.013	.748±.013	.567±.015	.613±.015	.708±.019	–	.112±.033	.229±.034
	Mistral-24B	.600±.012	.640±.013	.795±.009	.540±.020	.597±.019	.778±.013	–	.564±.021	.655±.026
	Meditron-7B	.422±.019	.463±.020	.709±.018	.336±.026	.399±.027	.696±.022	–	.142±.035	.141±.025

Table A16: Additional lexical evaluation for Task C timeline summarization. We report ROUGE-1 and ROUGE-2 F1 under SILVER→SILVER, SILVER→GOLD, and GOLD. Results are mean ± bootstrap SD over 2,000 document-level resamples. Higher is better.

Metric	Value	95% CI	Metric	Value	95% CI
Observed agreement (P_o)	0.904	[0.808, 0.981]	<i>Totals ($n = 52$ summaries)</i>		
Cohen’s κ	0.773	[0.567, 0.947]	Errors flagged by judge	262	–
Gwet’s AC1	0.835	[0.674, 0.966]	Errors upheld by annotator	39	–
Positive agreement	0.839	[0.667, 0.960]	Inflation ratio	6.7×	–
Negative agreement	0.932	[0.862, 0.986]	<i>Concordance</i>		
			Lin’s CCC	0.103	[0.031, 0.345]
			ICC(A,1)	0.105	[0.032, 0.349]
			Pearson’s r	0.511	[0.119, 0.818]
			Spearman’s ρ	0.768	[0.573, 0.934]
			<i>Bland–Altman analysis</i>		
			Bias (judge – annotator)	+4.29	SD = 15.32
			Limits of agreement	–25.7 to +34.3	

Table A17: Agreement between the automated evaluator and human annotations. P_o denotes observed agreement. Confidence intervals are reported at the 95% level.

M.1 Case A: Chronic Disability and an In-Stay Depression

Case A comprises 306 gold events across a 3-week admission. The record describes roughly three decades of C5 quadriplegia, an Individual-level risk factor that Stage 2 resolves to a year-only [APPROX] date some thirty years before admission, and a marked deterioration in ventilatory status during the stay. The note itself states the sequence, attributing the patient’s depressed mood to that deterioration, so the link is explicit rather than inferred by us. Depression is documented as a diagnosis: pharmacotherapy for situational depression is started during the admission, and psychiatry is consulted on the 11th hospital day. The symptoms recorded are suicidal thoughts and refusal of food; at discharge the note records restored hope, concentration and sleep, and continuation of

Table A18: Error-count concordance between the automated judge and human annotator. Per-summary agreement between the judge’s error count and the number of errors upheld by the annotator. CCC denotes Lin’s concordance correlation coefficient; ICC(A,1) denotes the absolute-agreement intraclass correlation coefficient.

an SSRI. Several symptoms are therefore attested only through their documented reversal, an implicit form of expression that an extraction system keyed to symptom mentions alone would miss.

What makes this case instructive is where the onset falls. The background risk factor begins decades before admission, whereas the depressive episode emerges during the hospital stay, making their temporal relationship difficult to recover from

Figure A3: The human review and correction platform. For each subject, the reviewer compares the model’s Stage 3 summary against the source and corrects it in place. Left: source panel, toggled between the discharge summary and Stage 2 gold events. Right: the model’s Stage 3 summary in an editable field, saved as a separate field. Patient-text regions from the original MIMIC III source materials are redacted (grey).

the narrative alone. Only a calendar-anchored representation places them on one ordered chain while preserving that separation. The distinction is also fragile in exactly the way our error analysis predicts: had the onset been tagged [PRE ADM] rather than resolved to the admission interval, the clinically salient fact that this depression arose during the stay would disappear from the timeline entirely. This is the same boundary that accounts for the largest class of temporal disagreements in Section G.

M.2 Case B: Depression Across All Three WHO Risk Levels

Case B contains 494 gold events spanning a 1.5-month admission after an intentional overdose. It is the only case in our sample with documented risk factors across all three WHO levels. At the individual level, the record describes long-standing alcohol misuse across the past medical history, social history, and discharge diagnoses. At the Family & Community level, it notes limited social support, poor self-care, and the recent loss of a living arrangement with a relative. At the Structural level, it identifies multiple social stressors and a lack of insurance coverage. These links are stated directly in the record, which cites them as reasons for in-

volving psychiatry and social work. The associated symptoms and behaviors include the presenting overdose, two earlier suicide attempts by different means, a discharge diagnosis noting multiple attempts, and a documented reduction in suicidal intent during the admission.

This case highlights two important features of the reconstruction. First, the structural risk factors are exactly the kind of information that a fixed clinical schema might omit, even though they are central to the explanation recorded by the care team. This illustrates the need for the open-vocabulary representation described in Section 1. Second, the record contains an unresolved contradiction: psychiatry documents that the patient was not currently depressed during the admission, while depression remains listed among the discharge diagnoses.

M.3 Case C: A Risk Cascade Under an Unsettled Diagnosis

Case C contains 227 gold events spanning a 12-day admission and includes four documented risk factors, only one of which the source explicitly links to the diagnosis. The record identifies more than two decades of alcohol misuse as clinically relevant to the patient’s diagnosis. It also documents childhood sexual abuse by a caregiver, job

loss attributed to drinking, and living alone after the end of a relationship about a year before admission. Because the note does not connect these latter factors directly to the disorder, we treat them as co-documented rather than causal. Bipolar disorder appears as an established diagnosis, supported by mood-stabilizing treatment at admission, repeated overdose-related hospitalizations, a prior psychiatric admission after a suicide attempt, and multiple presentations following suicide threats.

What distinguishes this case is that the record later revises its own clinical interpretation. During the admission, psychiatry questions whether the bipolar diagnosis is accurate and considers whether long-term alcohol use may have contributed to a misdiagnosis. Most background risk factors are undated and therefore tagged [PRE ADM], whereas this diagnostic reconsideration is placed within the admission interval. Temporal anchoring makes the progression visible: long-standing background factors, an established diagnosis, and then in-stay doubt about that diagnosis. The case also shows that uncertainty is not only temporal; it also concerns causality and diagnosis. Finally, it demonstrates the limits of surface matching. During screening, a statement about episodic drinking was incorrectly matched to an eating-disorder symptom term and was removed after clinician review, illustrating why lexical overlap alone cannot support reliable qualitative interpretation.

N Additional Results for Evaluation

We present additional evaluation results for the open-weight models (Table [A16](#)).