Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CliniCIRCA: LLM-powered Clinical Timeline Generation and Summarization from Unstructured EHR Texts for Mental Health Patients

Aiwei Ivy Zhang

azhang677@gatech.edu

Georgia Institute of Technology

Nimra Ishfaq

The University of Texas at Austin

Mohit Chandra

Georgia Institute of Technology

Santiago Alvarez Lesmes

Northwell Health

Adam Coscia

Stevens Institute of Technology

Khatiya Chelidze Moon

Northwell Health

Xiaohan Ding

Georgia Institute of Technology

Munmun De Choudhury

Georgia Institute of Technology

Research Article

Keywords: Natural language processing, Large language models, Electronic health records, Temporal reasoning, Mental health informatics, Clinical decision support, Information extraction, Medical text summarization

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11472328/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. De Choudhury serves on OpenAI's Expert Council on Well-Being and AI, on the Commission for Social Connection at the World Health Organization, on the advisory board of Australian eSafety Commission's Social Media Minimum Age Legislation, and as a Commissioner on Nature Health Commission on Consumer AI & Mental Wellbeing.

CliniCIRCA: LLM-powered Clinical Timeline Generation and Summarization from Unstructured EHR Texts for Mental Health Patients

Aiwei Ivy Zhang[1*], Nimra Ishfaq[2], Mohit Chandra[1], Santiago Alvarez Lesmes[3], Adam Coscia[4], Khatiya Chelidze Moon[3], Xiaohan Ding[1], Munmun De Choudhury[1]

[1]School of Interactive Computing, Georgia Institute of Technology, 85 5th St NW, Atlanta, 30309, GA, United States of America.
[2]School of Information, University of Texas at Austin, 1616 Guadalupe St Ste 5.202, Austin, 78701, Texas, United States of America.
[3]Northwell Health, 270-05 76th Ave, New Hyde Park, 11040, New York, United States of America.
[4]School of Computing, Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, 07030, New Jersey, United States of America.

*Corresponding author(s). E-mail(s): azhang677@gatech.edu;
Contributing authors: nimraishfaq@utexas.edu; mchandra9@gatech.edu; SALesmes@northwell.edu; acoscia@stevens.edu; KMoon2@northwell.edu; xding301@gatech.edu; munmund@gatech.edu;

Abstract

Background Mental health care often requires clinicians to reason over patient journeys: longitudinal patterns of symptoms, diagnoses, treatments, behaviors, social context, and responses to care. Yet these journeys are rarely documented as readily available, explicit timelines. Clinically important temporal information is often embedded in unstructured discharge narratives, where events may be

repeated, narrated out of order, or expressed through vague and relative time references. This makes temporal recovery difficult for clinicians and computational systems, and can limit the reliability of downstream summarization.

Methods We present `CliniCIRCA`, a multi-stage large language model (LLM) framework for **C**alendar-anchored, **I**mprecision-aware **R**econstruction of **C**linical **A**nnals. `CliniCIRCA` decomposes patient-journey reconstruction into three stages: extracting atomic clinical events from raw discharge summaries, assigning calendar-grounded temporal labels with explicit uncertainty, and rendering the verified timeline as a date-grouped clinical summary. The framework is designed to make temporal reasoning inspectable rather than hidden inside a single summarization step.

Results From 14,882 MIMIC-III mental-health-coded admissions, we construct a benchmark of 52 discharge summaries on which `CliniCIRCA` produces 15,891 temporally tagged events. Following clinician-defined rubrics, human annotation with clinician adjudication corrects 629 errors, producing verified gold-standard timelines. These corrected timelines then drive a temporally grounded summarization stage that compresses each source 1.52 times into a date-grouped chronological record while preserving temporal labels and uncertainty. We further scale the framework to generate 1,000 silver-standard timelines and evaluate them as supervision for smaller open-weight models. Compared with zero- and few-shot prompting, instruction tuning generally improves five open-weight models across event extraction, temporal tagging, and summarization under both silver and clinician-verified evaluations.

Conclusions Auditable temporal reconstruction can serve as both a clinician-readable representation of complex mental health narratives and a reusable supervision source for smaller models. These findings suggest that intermediate, temporally grounded representations may help bridge clinical summarization, timeline generation, and scalable model adaptation, while preserving the need for expert adjudication in high-stakes clinical natural language processing (NLP).

Keywords: Natural language processing, Large language models, Electronic health records, Temporal reasoning, Mental health informatics, Clinical decision support, Information extraction, Medical text summarization

1 Background

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 [1, 2]. Clinicians draw on this progression to interpret a patient's current presentation and guide treatment.

Unfortunately, these journeys are rarely documented as explicit timelines [3–5]. Much of a patient's history instead resides in free-text narratives spread across disparate sources [6, 7]. The burden of reconstruction therefore falls on clinicians, who must piece together the patient's course from a growing volume of documentation under time and information constraints [4, 5].

Discharge summaries are a particularly promising source for recovering these journeys. Written by clinicians at the end of a hospital stay, they synthesize prior history, hospital course, and clinical outcomes into a single narrative [8, 9]. Their temporal structure, however, is largely implicit. Events may be repeated, narrated out of sequence, or anchored only through relative expressions (e.g., "2 days prior to admission...", "5 years ago...") and surrounding context [10, 11], and temporal information varies in precision, from exact dates to vague references such as "remote history." These difficulties are compounded by abbreviations, typographical errors, fragmented syntax, and inconsistent formatting [12, 13], making the recovery of temporal structure a central challenge in clinical NLP [14, 15].

Automatic summarization with clinical timelines offers a natural way to reduce clinicians' reviewing burden, but a summary generated directly from clinical text inherits its implicit temporal structure and risks conflating past and present events, for example presenting a resolved condition as active [16]. For clinical review, a summary should not only condense a patient's record but also make clear when events occurred and group events under the same temporal bucket to show progression. This motivates first recovering an explicit, time-labeled representation of events, and then using it to organize the summary by time.

Prior work has largely treated this problem as separate subtasks: extracting temporal information, generating patient timelines, and summarizing clinical records. These lines of work rarely connect, and summaries are seldom generated from an explicit timeline as a direct way to preserve temporal cues.

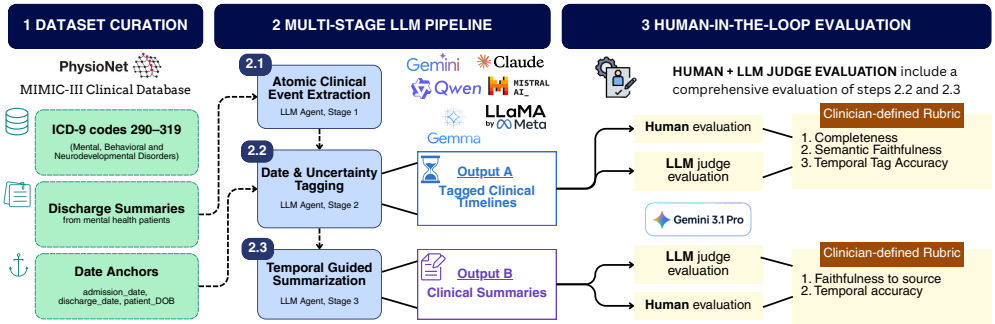


Fig. 1 Methodology Overview

Temporal information extraction recovers fine-grained structure by identifying time expressions, events, and temporal relations, but is typically built around predefined annotation schemas and pairwise relation classification rather than recovery of an open-ended patient trajectory [17–19]. Recent work also uses LLMs to generate event–time sequences through single-pass extraction or multi-stage reconstruction, coming closer to full patient trajectories [20–22]. These approaches make the task tractable by imposing structure that narrows their setting. Some restrict which events are captured: MedTimeline targets a predefined schema of chemotherapy events [23]. Some rely on structured EHR data to improve or calibrate event timing [20, 24]. Others limit patient-level context, either by chunking and retrieving fragments of discharge summaries [25] or by working with published case reports, which are more organized and curated than routine clinical narratives [21]. Time itself is commonly represented as a single relative offset [25], treating each event as precisely datable. Hence, events that are temporally vague or whose timestamp has uncertainty are either unmodeled [23] or forced into a precise timestamp [26].

Clinical summarization addresses the complementary goal of condensing lengthy records into coherent accounts for review and decision-making [9, 27, 28]. Most systems generate directly from the source note, combining event selection, temporal reasoning,

and narrative generation in a single step. As a result, temporality is often under-preserved, complex or uncertain chronology is rarely represented explicitly [26, 28, 29], and the chronology behind a summary is difficult to inspect.

What remains missing is a unified, text-only formulation that reconstructs a patient's timeline while preserving event breadth, calendar timing, temporal uncertainty, and patient-level coherence, and that uses this timeline directly to guide summarization. We address this gap with `CliniCIRCA`, which first constructs an event-level, uncertainty-aware timeline from discharge summaries and then summarizes from that representation, grouping events by temporal granularity. The summary is therefore not an independent compression of the original note, but a readable realization of an auditable timeline.

We study this problem in mental health care, where these requirements are especially pressing. Following Engel's biopsychosocial model [30], a psychiatric patient's course is shaped not only by diagnoses and procedures but also by medication use, trauma, social history, and changes in daily functioning [31]. These events are part of the trajectory rather than peripheral context, yet which ones matter differs across patients. Mental health discharge summaries therefore expose the limits of fixed schemas: a narrow schema omits events that may define an individual's course, while an exhaustive one quickly becomes impractical. This makes them a demanding testing ground for **open-vocabulary, calendar-anchored, and uncertainty-aware temporal reconstruction**.

`CliniCIRCA` (Figure 1) is a three-stage LLM framework that reconstructs each journey within a single discharge summary for one hospital admission, as a step toward multi-encounter longitudinal modeling. It produces two linked representations from the same temporal substrate: an inspectable event-level timeline for computational use and auditing (Output A), which then generates a clinician-readable, temporally grouped summary (Output B). Our experiments show that journeys produced by the

best-performing model in our pipeline (Gemini 2.5 Pro; Section 3) provide effective supervision for substantially smaller open-weight models, particularly for atomic clinical event extraction, while temporal reconstruction remains reasoning-intensive. By turning fragmented narratives into linked structured and readable representations, `CliniCIRCA` could support longitudinal chart review, clinical handoffs, and follow-up care while preserving an auditable link to the underlying events.

2 Methods

2.1 Dataset

We use MIMIC-III v1.4 [32, 33], 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. Dates within the discharge summaries are de-identified by shifting calendar dates at the patient level. Because the same offset is applied within a patient record, relative timing and within-patient event order are preserved, allowing admission, discharge, and birth dates to serve as temporal anchors while preventing interpretation of the shifted dates as real calendar dates. 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.

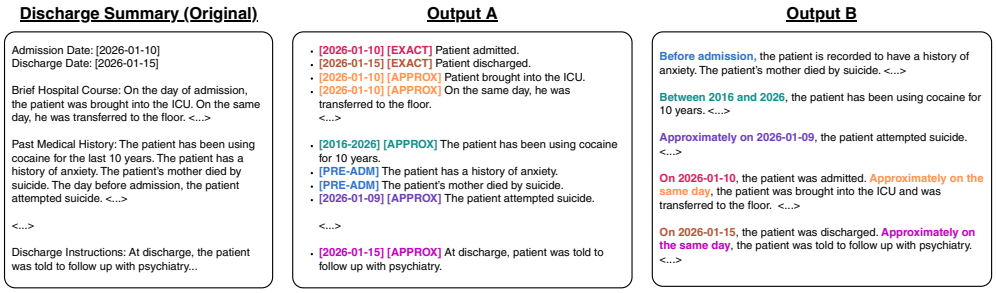


Fig. 2 Input (Original Discharge Summary) and Output (Output A & Output B) from left to right

2.2 The Clinical Timeline And Temporal Summarization Generation Framework

`CliniCIRCA` is a sequential framework comprising three LLM stages (Figure 2), each receiving 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.

2.3 Agent and Benchmark Setup

2.3.1 Agent Candidate Pool

For `CliniCIRCA`, we select 11 LLM candidates based on reported performance across general-purpose and clinical benchmarks [34–36]. For proprietary models, we evaluate Gemini 2.5 Pro and 2.5 Flash [37]; Claude Sonnet 4.6 [38] and Haiku 4.5 [39]. For open-weight models, we evaluate Llama 3.3 70B [40], Qwen 3.6 35B-A3B [41], Gemma 4 26B-A4B [42], Mistral Small 3.2 24B [43], Phi-4 Mini 3.8B [44], and medical-domain

LLMs - MedGemma 27B [45] and Meditron 3 Qwen 2.5 7B [46]. Model specifications, generation settings, and computational details are provided in Additional File 1.

2.3.2 Benchmark Implementation Setup

We construct a benchmark set of 52 MIMIC-III discharge summaries, averaging 1,774 words each (range: 437–3,583). 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 (Additional File 1) for a pairwise χ^2 comparison (large effect, $\alpha = 0.05$, 95% power), yielding $N = 52$ needed to reach statistical significance [47–49].

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 2.4. 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 `CliniCIRCA` framework, and simplifies reproducibility and deployment governance [50–52]. 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 [53], we use a clinician-defined evaluation process in which clinicians specify the rubrics and error-counting guidelines, two human annotators apply these guidelines, and clinicians adjudicate disagreements and ambiguous cases. All annotators and clinicians have prior data annotation experience, and are proficient in English. A calibrated LLM judge independently scores the same outputs; after validation against human annotations and clinician-adjudicated labels, it is used to scale Stage-2 model selection. We establish human-evaluator agreement on both Stage-2 and 3 outputs.

2.4 Stage 1 & 2 Tagged Timeline Generation

Faithful clinical timeline recovery requires both comprehensive event extraction and accurate temporal anchoring, yet extraction errors can propagate through temporal reasoning and timeline construction [18, 19]. 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 Additional File 2 Tables.

2.4.1 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.

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.

2.4.2 Evaluation Design for Stages 1 & 2

Because no reference timeline exists against which these tuples could be scored, a "good" timeline must itself be operationalized. Clinicians defined three criteria for evaluating the Stage-2 output, including task-specific guidelines for counting errors and

assigning error types for evaluating the Stage-2 output (full definitions and evaluation platform in Additional File 3).

A) Completeness: *whether all clinical events, including all events related to the patient's mental health, 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 temporal certainty tag(s) are correct.*

Following established practice in LLM evaluation to use a strong later-generation model [54], 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. Human review was conducted on an internally hosted, secure annotation platform, where annotators reviewed each generated timeline against the source discharge summary and recorded criterion-specific errors and comments.

2.5 Stage 3 Temporal Summarization Generation

Next, Stage 3 tests if the verified event timeline can be rendered as a faithful, temporally grounded summary. The generation prompt can be found in Additional File 4

2.5.1 Temporal-Guided Summaries using Outputs from Stage 2

Two clinicians defined a mental-health-oriented summarization specification that retains demographics and social history, reason for admission, key in-stay events, medication changes, discharge, and follow-up while preserving the temporal labels assigned in Stage 2.

A) Faithfulness to Source: *whether the summary preserves clinical information from the verified timeline without adding unsupported claims or distorting event*

meaning.

B) Temporal Accuracy: *whether the summary preserves the dates, temporal labels, and uncertainty of retained events from the verified timeline.*

2.5.2 Evaluation Design for Stage 3

Full rubric definitions and the human evaluation platform are available in Additional File 5. As in the Stage 1–2 evaluation, clinicians defined the Stage-3 rubric and error-counting guidelines, and two human annotators applied them to the generated summaries. Each Stage-3 summary required 18 minutes to review on average. Disagreements and ambiguous cases were reviewed in additional clinician-led adjudication sessions, where clinicians provided the final gold labels.

2.6 Downstream Instruction Tuning for Small Open-weight Models

We evaluate whether `CliniCIRCA`'s intermediate outputs can serve as reusable supervision under three training and evaluation settings.

Silver→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 (Additional File 6).

Silver→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.

2.6.1 Experiment 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 specialization: Gemma4-31B [42], Mistral-24B [43], Qwen3.6-35B-A3B [41], Qwen3-30B-A3B [55], and Meditron-7B [46]. 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 [56, 57]. 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 which are additionally reported in Additional File 7.

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.

3 Results

3.1 Clinician-in-the-Loop Evaluation Results for Generation

3.1.1 Stage 1 & 2 Tagged Clinical Timelines

Best Model for Clinical Timeline Generation and Summarization

Following the agent selection mechanism as described in Section 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 [58], 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.

Table 1 Stage-2 model-selection results on the 52-sample benchmark.

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**	4.98	4.81	4.93	0.610	0.753	0.672
Claude Sonnet 4.6	4.98	**5.00**	4.77	4.92	0.614	0.735	0.666
Gemini 2.5 Flash	4.98	**5.00**	4.50	4.83	0.617	0.762	0.679
Qwen 3.6 35B-A3B	**5.00**	**5.00**	4.23	4.74	0.630	0.758	0.685
Gemma 4 26B-A4B	4.88	4.96	4.27	4.71	0.636	0.713	0.669
Mistral Small 3.2 24B	4.96	4.98	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	4.98	4.98	2.67	4.21	0.638	0.729	0.678
Meditron 3 (Qwen 2.5 7B)	3.71	4.73	2.00	3.48	0.680	0.285	0.361
Phi-4 Mini 3.8B	3.90	4.17	2.19	3.42	**0.685**	0.461	0.525

Results are averaged over 52 samples. Dimension A denotes Completeness, Dimension B denotes Semantic Faithfulness, and Dimension C denotes Temporal Tag Accuracy. Dimensions A–C are Likert-scale averages reported to two decimal places. The bold model name in the first row identifies the model selected for the final pipeline. For each metric, the best-performing value is shown in **bold** and the second-best value is underlined. Tied values receive the same formatting.

Table 2 Reading-burden reduction from Stage 3 summarization over the 52-note benchmark.

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

Per-note percentages are averaged across notes; corpus-level reduction pools word counts across all notes.

Agreement and Error Analysis for the Stage 1-2 Timeline

We compare the calibrated Gemini 3.1 Pro evaluator with the clinician-adjudicated gold labels produced from human annotation (see Additional File 3 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 [59]. 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. We provide additional analysis of agreement metrics for Stage 2 outputs in Additional File 8. Overall, the LLM judge was suitable for scaling model selection, but clinician adjudication remained necessary for constructing the final reference labels.

3.1.2 Stage 3 Clinical Summaries

Agreement and Error Analysis for the Stage 3 Temporal Summaries

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 Additional File 4 to 5 and results are demonstrated in Table 3. Across the benchmark, Stage 3 reduces reading burden

Table 3 Evaluation results for summaries generated by Gemini 2.5 Pro.

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

Faithfulness and temporal accuracy are reported as mean rubric bands and percentage scores. MEDCON precision, recall, and F1 measure clinical-concept overlap.

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).

We compare the judge's error counts against human adjudication across 52 summaries. 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 $\rho = 0.77$). See Additional File 9 for full agreement and error analysis.

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 [60], and three levels of risk factors: Individual, Family & Community, and Structural [61]. 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 Additional File 10.

3.2 Baseline and Ablation Comparisons

A direct baseline comparison to prior clinical timeline systems is not straightforward because existing methods differ from CliniCIRCA in input format, prompting assumptions, temporal representation, and output structure. In particular, CliniCIRCA reconstructs open-vocabulary timelines directly from raw discharge summaries and explicitly preserves temporal uncertainty through `[EXACT]`, `[APPROX]`, `[PRE ADM]`, and `[INDETERMINATE]` labels, which are then used as the substrate for Stage 3 summarization. Prior systems that often require every event to receive a single relative timestamp cannot represent these distinctions without forcing uncertain, historical, or indeterminate events into a more precise temporal form than the source supports.

3.2.1 Within-Framework Baseline

As a within-framework baseline, we also evaluated single-stage variants that collapse CliniCIRCA's modular decomposition, with more details in Additional File 11. For the timeline task up to Stage 2, one prompt jointly performed event extraction and temporal tagging; for the summarization task up to Stage 3, one prompt generated the final summary without using the verified intermediate timeline as a separate substrate. This comparison keeps the dataset, model family, and evaluation criteria aligned while testing whether decomposition itself is useful. The single-stage variants produced lower LLM-judge scores and introduced human-identified errors that were reduced by the staged pipeline, supporting the design choice to separate event extraction, temporal anchoring, and temporally guided summarization, especially for long and noisy discharge summaries.

3.2.2 Cross Comparison

We evaluated the released prompting pipeline of Wang and Weiss [22, 62] on our 52-note benchmark, using Gemini 2.5 Pro in place of GPT-4-0613. Both systems were

evaluated by Gemini 3.1 Pro using the same rubrics (Additional File 12). Wang and Weiss's formulation represents each event using a signed offset in hours from admission, whereas CliniCIRCA explicitly represents dates and temporal uncertainty. The reported completeness rates were 92.0% for the Wang and Weiss formulation and 99.9% for CliniCIRCA; semantic faithfulness rates were 94.8% and 99.9%, respectively. Temporal-tag accuracy was 73.4% and 97.7%, respectively. Interpretation of the temporal comparison requires caution because the output representations are different. To assess the baseline's temporal tags, the judge converted relative offsets into calendar dates and supplied temporal labels that were absent from the original output. The resulting scores therefore depend on this conversion and on compatibility with our evaluation rubric. We treat this analysis as a diagnostic comparison, with detailed results and limitations provided in Additional File 12.

3.3 Results on Instruction Tuning Small Open-weight Models

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 Table 4. 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 clinical 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 4-Task A). 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

Table 4 Open-weight model evaluation across Tasks A–C.

Task A: Atomic Clinical Event Extraction

	Silver→Silver			Silver→Gold			Gold		
Model	**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	$.536_{.014}$	$.642_{.013}$	$.839_{.009}$	$.535_{.021}$	$.647_{.017}$	$.834_{.012}$	–	$\mathbf{.732}_{.012}$	$.550_{.026}$
Qwen3-30B-A3B	$.420_{.027}$	$.547_{.025}$	$.745_{.028}$	$.454_{.028}$	$.561_{.020}$	$.795_{.013}$	–	$.279_{.045}$	$.606_{.033}$
Mistral-24B	$.309_{.021}$	$.565_{.016}$	$\mathbf{.848}_{.008}$	$.284_{.021}$	$.545_{.032}$	$\mathbf{.847}_{.009}$	–	$.676_{.015}$	$.715_{.028}$
Meditron-7B	$.040_{.006}$	$.131_{.013}$	$.744_{.021}$	$.032_{.007}$	$.120_{.015}$	$.752_{.021}$	–	$.034_{.004}$	$.331_{.024}$

Task B: Temporal Tagging

	Silver→Silver			Silver→Gold			Gold		
Model	**Zero**	**Few**	**IT**	**Zero**	**Few**	**IT**	**Zero**	**Few**	**IT**
Gemma4-31B	$.738_{.019}$	$.731_{.019}$	$\mathbf{.767}_{.024}$	$.799_{.036}$	$\mathbf{.804}_{.043}$	$.749_{.029}$	–	$.668_{.029}$	$\mathbf{.721}_{.035}$
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	$.266_{.016}$	$.004_{.003}$	$.401_{.035}$	$.256_{.018}$	$.004_{.004}$	$.666_{.048}$	–	$.005_{.004}$	$.359_{.022}$

Task C: Clinical Timeline Summarization

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

Legend. Task A is measured by embedding-based RM-F1@0.85, Task B by macro-averaged F1 over the four temporal tags, and Task C by ROUGE-L. Results are point estimates with bootstrap SD shown as subscripts.

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-Task B). 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 4-Task C). 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

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.

4 Discussion

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 [63–65]. `CliniCIRCA` 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 1).

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 [66, 67]. 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 [68, 69]. 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 [70–72].

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 [73]. 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.

4.1 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 2, the results may understate disagreement in salience or interpretation across a broader clinician population. 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 (Additional File 4) 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 3.1.2 – 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.

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

5 Conclusions

`CliniCIRCA` 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 `CliniCIRCA` to multi-document patient records, richer temporal representations, and additional note types, to support traceable longitudinal chart review and memory-aware clinical systems.

6 List of Abbreviations

LLM: Large language model

NLP: Natural language processing

EHR: Electronic Health Record

HTML: Hypertext Markup Language

ED: Emergency department

MoE: Mixture of experts

LoRA: Low-Rank Adaptation

API: Application programming interface

WHO: World Health Organization

UMLS: Unified Medical Language System

Additional Files

Additional File 1.

Pipeline Agent Candidate & Benchmark Size Selection

File format: PDF (.pdf)

Description: Details of pipeline agent candidate selection and benchmark size determination.

Additional File 2.

Stage 1–2 Generation Prompts

File format: PDF (.pdf)

Description: Prompt templates used to generate outputs in Stages 1 and 2.

Additional File 3.

Stage 1–2 Evaluation Rubrics and Platform

File format: PDF (.pdf)

Description: Evaluation rubrics and the platform used to assess outputs from Stages 1 and 2.

Additional File 4.

Stage 3 Generation Prompts

File format: PDF (.pdf)

Description: Prompt templates used to generate summaries in Stage 3.

Additional File 5.

Stage 3 Evaluation Rubrics and Platform

File format: PDF (.pdf)

Description: Evaluation rubrics and the platform used to assess Stage 3 summaries.

Additional File 6.

Token Usage for Silver-Standard Supervision

File format: PDF (.pdf)

Description: Token usage associated with generating silver-standard supervision data.

Additional File 7.

Additional Results for Stage 3 Instruction Tuning on Smaller Open Models

File format: PDF (.pdf)

Description: Supplementary results for Stage 3 instruction tuning using smaller open models.

Additional File 8.

Agreement & Error Analysis for Stage 2

File format: PDF (.pdf)

Description: Agreement assessment and error analysis for Stage 2 outputs.

Additional File 9.

Agreement & Error Analysis for Stage 3

File format: PDF (.pdf)

Description: Agreement assessment and error analysis for Stage 3 summaries.

Additional File 10.

Qualitative Analysis of Three Stage 3 Summaries

File format: PDF (.pdf)

Description: Qualitative examination of three Stage 3 summaries, highlighting their strengths and limitations.

Additional File 11.

Within-framework Baseline Study

File format: PDF (.pdf)

Description: Methods and results of baseline comparisons within the proposed framework.

Additional File 12.

Cross Comparison Baseline Study

File format: PDF (.pdf)

Description: Methods and results of the cross-comparison baseline study.

Declarations

Ethics approval and consent to participate. 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 – our pipeline'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.

Consent for publication. Not applicable.

Availability of data and materials. The study uses MIMIC-III Clinical Database v1.4, a restricted-access de-identified dataset available through PhysioNet to credentialed users who complete the required data-use agreement and training. The derived

patient-level outputs cannot be publicly released under the data-use agreement. We provide code in the following GitHub repository, where all examples are synthetic: https://github.com/theivyzhang/CliniCIRCA.

Competing interests. De Choudhury serves on OpenAI's Expert Council on Well-Being and AI, on the Commission for Social Connection at the World Health Organization, on the advisory board of Australian eSafety Commission's Social Media Minimum Age Legislation, and as a Commissioner on Nature Health Commission on Consumer AI & Mental Wellbeing.

Funding. Zhang, Ding, and De Choudhury were partly supported through funds from The Children's Healthcare of Atlanta Pediatric Technology Center at Georgia Tech.

Authors' contributions. AIZ conceived and led the study, including project design, data curation and analysis, pipeline development, experimental design and execution, evaluation design, downstream modeling experiments, results interpretation, and manuscript drafting and revision. NI contributed to study conceptualization and design, as well as guidance for clinician-in-the-loop evaluation and expert annotations. MC contributed to study conceptualization and design, and technical advising with LLM generation. SAL contributed to clinician adjudication of generated outputs and interpretation of evaluation results. AC contributed to initial study scope design and dataset selection. KCM contributed to and supervised medical concept validation, pipeline design, clinician adjudication of generated outputs, and interpretation of evaluation results. XD supervised task execution, with particular focus on framework design, model fine-tuning, and distillation studies. MDC contributed to and supervised study conceptualization and design, result interpretation, and manuscript writing. All authors read and approved the manuscript.

Acknowledgements. 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

[1] Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. The Lancet Digital Health. 2024 Apr;6(4):e281–e290. https://doi.org/10.1016/S2589-7500(24)00025-6.

[2] Ruan T, Lei L, Zhou Y, Zhai J, Zhang L, He P, et al. Representation learning for clinical time series prediction tasks in electronic health records. BMC Medical Informatics and Decision Making. 2019 Dec;19(S8):259. https://doi.org/10.1186/s12911-019-0985-7.

[3] Loftus TJ, Balch JA, Marquard JL, Ray JM, Alper BS, Ojha N, et al. Longitudinal clinical decision support for assessing decisions over time: State-of-the-art and future directions. DIGITAL HEALTH. 2024 Jan;10:20552076241249925. https://doi.org/10.1177/20552076241249925.

[4] Linhares CDG, Lima DM, Ponciano JR, Olivatto MM, Gutierrez MA, Poco J, et al. ClinicalPath: A Visualization Tool to Improve the Evaluation of Electronic Health Records in Clinical Decision-Making. IEEE Transactions on Visualization and Computer Graphics. 2023 Oct;29(10):4031–4046. https://doi.org/10.1109/

TVCG.2022.3175626.

[5] Olex AL, McInnes BT. Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be. Journal of Biomedical Informatics. 2021 Jun;118:103784. https://doi.org/10.1016/j.jbi.2021.103784.

[6] Chaturvedi R. Temporal Knowledge Graph Extraction and Modeling across Multiple Documents for Health Risk Prediction. In: Companion Proceedings of the ACM Web Conference 2024. Singapore Singapore: ACM; 2024. p. 1182–1185. Available from: https://dl.acm.org/doi/10.1145/3589335.3651256.

[7] Moharasan G, Ho TB. Extraction of Temporal Information from Clinical Narratives. Journal of Healthcare Informatics Research. 2019 Jun;3(2):220–244. https://doi.org/10.1007/s41666-019-00049-0.

[8] Seinen TM, Kors JA, Van Mulligen EM, Rijnbeek PR. Using Structured Codes and Free-Text Notes to Measure Information Complementarity in Electronic Health Records: Feasibility and Validation Study. Journal of Medical Internet Research. 2025 Feb;27:e66910. https://doi.org/10.2196/66910.

[9] Lee C, Vogt KA, Kumar S. Prospects for AI clinical summarization to reduce the burden of patient chart review. Frontiers in Digital Health. 2024 Nov;6:1475092. https://doi.org/10.3389/fdgth.2024.1475092.

[10] K K, Thirukovalluru R, Carlson D. 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. Mumbai,

India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics; 2025. p. 2822–2836. Available from: https://aclanthology.org/2025.ijcnlp-long.151.

[11] Kim MK, Rouphael C, McMichael J, Welch N, Dasarathy S. Challenges in and Opportunities for Electronic Health Record-Based Data Analysis and Interpretation. Gut and Liver. 2024 Mar;18(2):201–208. https://doi.org/10.5009/gnl230272.

[12] Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdisciplinary Reviews: Computational Statistics. 2021;13. https://doi.org/10.1002/wics.1549.

[13] Frings J, Rust P, Jede F, Meister S, Prinz C, Fehring L. Diagnosis decoded: a taxonomy and natural language processing analysis of the diagnosis section in German hospital discharge summaries. Health Care Management Science. 2025;28:806 – 823. https://doi.org/10.1007/s10729-025-09732-8.

[14] Amirahmadi A, Etminani F, Björk J, Melander O, Ohlsson M. Trajectory-Ordered Objectives for Self-Supervised Representation Learning of Temporal Healthcare Data Using Transformers: Model Development and Evaluation Study. JMIR Medical Informatics. 2025 Jun;13:e68138. https://doi.org/10.2196/68138.

[15] Makarov N, Bordukova M, Quengdaeng P, Garger D, Rodriguez-Esteban R, Schmich F, et al. Large language models forecast patient health trajectories enabling digital twins. npj Digital Medicine. 2025 Oct;8(1):588. https://doi.org/10.1038/s41746-025-02004-3.

[16] Gao Y, Dligach D, Miller T, Xu D, Churpek MMM, Afshar M. Summarizing Patients' Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models. In: Calzolari N, Huang CR, Kim H, Pustejovsky J, Wanner L, Choi KS, et al., editors. Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea: International Committee on Computational Linguistics; 2022. p. 2979–2991. Available from: https://aclanthology.org/2022.coling-1.264/.

[17] Chaturvedi R, Baghershahi P, Medya S, Di Eugenio B. Temporal Relation Extraction in Clinical Texts: A Span-based Graph Transformer Approach. In: Che W, Nabende J, Shutova E, Pilehvar MT, editors. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics; 2025. p. 25765–25788. Available from: https://aclanthology.org/2025.acl-long.1251/.

[18] Gumiel YB, Silva E Oliveira LE, Claveau V, Grabar N, Paraiso EC, Moro C, et al. Temporal Relation Extraction in Clinical Texts: A Systematic Review. ACM Computing Surveys. 2022 Sep;54(7):1–36. https://doi.org/10.1145/3462475.

[19] Alfattni G, Peek N, Nenadic G. Extraction of temporal relations from clinical free text: A systematic review of current approaches. Journal of Biomedical Informatics. 2020 Aug;108:103488. https://doi.org/10.1016/j.jbi.2020.103488.

[20] Kumar S, Noroozizadeh S, Kim J, Weiss JC.: Text Knows What, Tables Know When: Clinical Timeline Reconstruction via Retrieval-Augmented Multimodal Alignment. arXiv. Version Number: 1. Available from: https://arxiv.org/abs/2605.15168.

[21] Noroozizadeh S, Kumar S, Chen GH, Weiss JC.: Temporally annotated textual time series from PubMed Open Access clinical case reports. Health Informatics.

Available from: http://medrxiv.org/lookup/doi/10.1101/2025.11.01.25339297.

[22] Wang J, Weiss JC. 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;2025:598–606.

[23] Wang L, Lu Q, Li R, Fu S, Liu H. 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. Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 483–487. Available from: https://aclanthology.org/2024.clinicalnlp-1.48.

[24] Frattallone-Llado G, Kim J, Cheng C, Salazar D, Edakalavan S, Weiss JC. Using Multimodal Data to Improve Precision of Inpatient Event Timelines. In: Yang DN, Xie X, Tseng VS, Pei J, Huang JW, Lin JCW, editors. Advances in Knowledge Discovery and Data Mining. vol. 14648. Singapore: Springer Nature Singapore; 2024. p. 322–334. Series Title: Lecture Notes in Computer Science. Available from: https://link.springer.com/10.1007/978-981-97-2238-9_25.

[25] Wang J, Niu X, Kim J, Shen J, Zhang T, Weiss J.: MIMIC-IV-Ext-22MCTS: A 22 Millions-Event Temporal Clinical Time-Series Dataset with Relative Timestamp for Risk Prediction. In Review. Available from: https://www.researchsquare.com/article/rs-6347897/v1.

[26] Kruse M, Hu S, Derby N, Wu Y, Stonbraker S, Yao B, et al. 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;2025:20715 – 20735. https://doi.org/10.18653/v1/2025.findings-emnlp.1128.

[27] Luo Z, Ji Y, Gupta A, Li Z, Frisch A, He D. 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). Pittsburgh, PA, USA: IEEE; 2023. p. 1–5. Available from: https://ieeexplore.ieee.org/document/10313411/.

[28] Croxford E, Gao Y, Pellegrino N, Wong K, Wills G, First E, et al. Development and validation of the provider documentation summarization quality instrument for large language models. Journal of the American Medical Informatics Association. 2025 Jun;32(6):1050–1060. https://doi.org/10.1093/jamia/ocaf068.

[29] Cui H, Unell A, Chen B, Fries JA, Alsentzer E, Koyejo S, et al. TIMER: temporal instruction modeling and evaluation for longitudinal clinical records. npj Digital Medicine. 2025 Sep;8(1):577. https://doi.org/10.1038/s41746-025-01965-9.

[30] Engel GL. The Need for a New Medical Model: A Challenge for Biomedicine. Science. 1977 Apr;196(4286):129–136. https://doi.org/10.1126/science.847460.

[31] Saha K, Yoo DW, Das Swain V, De Choudhury M. Life events as predictors of wellbeing outcomes. npj Digital Public Health. 2026 Mar;1(1):5. https://doi.org/10.1038/s44482-025-00005-3.

[32] Johnson AEW, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016 May;3(1):160035. https://doi.org/10.1038/sdata.2016.35.

[33] Johnson A, Pollard T, Mark R.: MIMIC-III Clinical Database. PhysioNet. Available from: https://physionet.org/content/mimiciii/1.4/.

[34] Yu H, Zhou J, Li L, Chen S, Gallifant J, Shi A, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming

medical education. Communications Medicine. 2025 Dec;6(1):27. https://doi.org/10.1038/s43856-025-01283-x.

[35] Sarvari P, Al-fagih Z. Rapidly Benchmarking Large Language Models for Diagnosing Comorbid Patients: Comparative Study Leveraging the LLM-as-a-Judge Method. JMIRx Med. 2025 Aug;6:e67661–e67661. https://doi.org/10.2196/67661.

[36] Shi W, Xu R, Zhuang Y, Yu Y, Zhang J, Wu H, et al. 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. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 22315–22339. Available from: https://aclanthology.org/2024.emnlp-main.1245.

[37] Gemini Team. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv preprint arXiv:250706261. 2025;https://doi.org/10.48550/arXiv.2507.06261.

[38] Anthropic.: Claude Sonnet 4.6 System Card. System card. Available from: https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf.

[39] Anthropic.: Claude Haiku 4.5 System Card. System card. Available from: https://www-cdn.anthropic.com/7aad69bf12627d42234e01ee7c36305dc2f6a970.pdf.

[40] Meta AI.: Llama 3.3 70B Instruct Model Card. Accessed July 27, 2026. Hugging Face model card. Available from: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

[41] Qwen Team.: Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All. Available from: https://qwen.ai/blog?id=qwen3.6-35b-a3b.

[42] Gemma Team.: Gemma 4 Technical Report. Available from: https://arxiv.org/abs/2607.02770.

[43] Mistral AI.: Mistral-Small-3.2-24B-Instruct-2506 Model Card. Accessed July 27, 2026. Hugging Face model card. Available from: https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506.

[44] Abouelenin A, et al. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv preprint arXiv:250301743. 2025;.

[45] Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al.: MedGemma Technical Report. Available from: https://arxiv.org/abs/2507.05201.

[46] OpenMeditron.: Meditron3-Qwen2.5-7B Model Card. Accessed July 27, 2026. Hugging Face model card. Available from: https://huggingface.co/EPFLiGHT/Meditron3-Qwen2.5-7B.

[47] Ki D, Carpuat M, McNamee P, Khashabi D, Yang E, Lawrie D, et al. Linguistic Nepotism: Trading-off Quality for Language Preference in Multilingual RAG. arXiv preprint arXiv:250913930. 2025;.

[48] Card D, Henderson P, Khandelwal U, Jia R, Mahowald K, Jurafsky D. With Little Power Comes Great Responsibility. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 9263–9274. Available from: https://aclanthology.org/2020.emnlp-main.745.

[49] Dror R, Baumer G, Shlomov S, Reichart R. 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). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 1383–1392. Available from: http://aclweb.org/anthology/P18-1128.

[50] Umeton R, Kwok A, Maurya R, Leco D, Lenane N, Willcox J, et al. GPT-4 in a Cancer Center — Institute-Wide Deployment Challenges and Lessons Learned. NEJM AI. 2024 Mar;1(4). https://doi.org/10.1056/AIcs2300191.

[51] de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KG, et al. From text to treatment: the crucial role of validation for generative large language models in health care. The Lancet Digital Health. 2024 Jul;6(7):e441–e443. https://doi.org/10.1016/S2589-7500(24)00111-0.

[52] Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digital Medicine. 2023 Jul;6(1):120. https://doi.org/10.1038/s41746-023-00873-0.

[53] Mittal S, Kasson E, Paraboschi L, Laufenberg E, Zhou J, Cavazos-Rehg PA, et al. 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; 2026. To appear. Available from: https://arxiv.org/abs/2603.19093.

[54] Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. 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.; 2023. .

[55] Qwen Team.: Qwen3 Technical Report. Available from: https://arxiv.org/abs/2505.09388.

[56] Sharif O, Gatto J, Basak M, Preum SM. REGen: A Reliable Evaluation Framework for Generative Event Argument Extraction. In: Findings of the Association for Computational Linguistics: EMNLP 2025; 2025. p. 12146–12168.

[57] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:190409675. 2019;.

[58] wai Yim W, Fu Y, Ben Abacha A, Snider N, Lin T, Yetisgen M. ACI-BENCH: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. Nature Scientific Data. 2023;.

[59] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology. 2008;61(1):29–48. https://doi.org/10.1348/000711006X126600.

[60] World Health Organization.: Mental Disorders. Accessed: 2026-07-13. https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[61] World Health Organization.: Mental Health. Accessed: 2026-07-14. https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response.

[62] Weiss JC.: LLM-Timeline-PMOA: A LLM Framework for Relative Timeline Extraction from PubMed Case Reports. Accessed 2026-10-21. GitHub repository. Available from: https://github.com/jcweiss2/LLM-Timeline-PMOA.

[63] Balasubramanian JB, Adams D, Roxanis I, De Gonzalez AB, Coulson P, Almeida JS, et al. Leveraging large language models for structured information extraction from pathology reports. Journal of Pathology Informatics. 2025 Nov;19:100521. https://doi.org/10.1016/j.jpi.2025.100521.

[64] Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, Papadopoulos N, Odavic D, Risteski P, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health & Care Informatics. 2025 Jan;32(1):e101139. https://doi.org/10.1136/bmjhci-2024-101139.

[65] Grothey B, Odenkirchen J, Brkic A, Schömig-Markiefka B, Quaas A, Büttner R, et al. Comprehensive testing of large language models for extraction of structured data in pathology. Communications Medicine. 2025 Mar;5(1):96. https://doi.org/10.1038/s43856-025-00808-8.

[66] Jiang Z, Chen H, Wu Y, Qin Y, Pei C, Zeng D, et al. Beyond multiple-choice questions: Rethinking evaluation frameworks for large language models for clinical medicine. Intelligent Medicine. 2026 Apr;6(2):109–115. https://doi.org/10.1016/j.imed.2026.01.001.

[67] Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine. 2024 Sep;30(9):2613–2622. https://doi.org/10.1038/s41591-024-03097-1.

[68] Chandra M, Sriraman S, Verma G, Khanuja HS, Campayo JS, Li Z, et al. 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). Albuquerque, New Mexico: Association for Computational Linguistics; 2025. p. 11083–11113. Available from: https://aclanthology.org/2025.naacl-long.553.

[69] Wang X, Zhang J, Zhang G, Guo H. Feel the Difference? A Comparative Analysis of Emotional Arcs in Real and LLM-Generated CBT Sessions. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V, editors. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics; 2025. p. 19999–20017. Available from: https://aclanthology.org/2025.findings-emnlp.1089/.

[70] Bavaresco A, Bernardi R, Bertolazzi L, Elliott D, Fernández R, Gatt A, et al. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In: Che W, Nabende J, Shutova E, Pilehvar MT, editors. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vienna, Austria: Association for Computational Linguistics; 2025. p. 238–255. Available from: https://aclanthology.org/2025.acl-short.20/.

[71] Diekmann Y, Fensore C, Carrillo-Larco R, Castejon Rosales E, Shiromani S, Pai R, et al. LLMs as Medical Safety Judges: Evaluating Alignment with Human Annotation in Patient-Facing QA. In: Proceedings of the 24th Workshop on Biomedical Language Processing. Viena, Austria: Association for Computational Linguistics; 2025. p. 217–224. Available from: https://aclanthology.org/2025.bionlp-1.19.

[72] Chen GH, Chen S, Liu Z, Jiang F, Wang B. Humans or LLMs as the Judge? A Study on Judgement Bias. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 8301–8327. Available from: https://aclanthology.org/2024.emnlp-main.474.

[73] Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine. 2024 Apr;30(4):1134–1142. https://doi.org/10.1038/s41591-024-02855-5.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile6.pdf
- AdditionalFile2.pdf
- AdditionalFile10.pdf
- AdditionalFile7.pdf
- AdditionalFile11.pdf
- AdditionalFile1.pdf
- AdditionalFile9.pdf
- AdditionalFile8.pdf
- AdditionalFile4.pdf
- AdditionalFile12.pdf
- AdditionalFile3.pdf
- AdditionalFile5.pdf