Additional File 12

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. MedTimeline targets a fixed chemotherapy schema; Frattallone-Llado et al. require structured EHR timestamps alongside the text; MIMIC-IV-Ext-22MCTS uses relative timestamps because MIMIC-IV-Note replaces in-text dates with placeholders, so calendar anchoring is foreclosed by that corpus rather than chosen against. A score computed across these settings would largely measure the mismatch introduced by forcing them together.

We nonetheless ran the closest published, peer-reviewed, purely prompt-based comparator end-to-end on our 52-note benchmark: Wang and Weiss (1), which emits one row per event as "event | t", a signed offset in hours from admission. We used their released *base_prompt* verbatim (2), extracted programmatically from their repository and asserted byte-identical, together with their message assembly and their output parser. The only forced deviation is the generation model (Gemini 2.5 Pro rather than GPT-4-0613), since MIMIC-III may only be sent to endpoints covered by our data use agreement. Both systems were then scored by the same judge (Gemini 3.1 Pro) under the same unmodified rubrics.

Cross-Comparison Baseline Study

Tasks for Stage 1 + Stage 2: Producing a temporally tagged timeline

From a technical standpoint, their output space cannot represent 14.3% of what our reference asserts: a signed offset always resolves to a calendar point, so the 2,238 [PRE ADM] and 28 [INDETERMINATE] events (14.1% and 0.2% of 15,891) have no representation in it, and the 61.0% of events labelled [APPROX] are expressible only by adding precision the source does not support. Under the shared evaluation rubrics, the largest difference was in temporal-tag accuracy: 73.4% for the Wang and Weiss formulation versus 97.7% for CliniCIRCA. Completeness rates were 92.0% versus 99.9%, and semantic faithfulness rates were 94.8% versus 99.9%, respectively.

Table S12.1.

Dimension	W&W band	W&W rate	CliniCIRCA band	CliniCIRCA rate
A Completeness	4.37	92.0%	5.00	99.9%
B Semantic faithfulness	4.67	94.8%	5.00	99.9%
C Temporal tag accuracy	3.44	73.4%	4.83	97.7%

Timeline extraction and temporal-tagging results.

Interpretation of these results depends on the compatibility of each output representation with the evaluation rubrics. We therefore emphasize results for individual evaluation dimensions. Because their rows carry no date and no tag, our judge had to supply both before grading, and it did so silently: of 1,747 temporal findings citing a specific date, only 6 referred to a row that actually contained one. The judge computed admission + t and graded its own conversion. We therefore treat this as a diagnostic comparison rather than a primary benchmark.

Tasks for Stage 3: Producing a temporally guided summary

Setup. Wang and Weiss's framework ends at timeline extraction, as it has no further summarization stage. To evaluate Stage 3 at all, we therefore took the MIMIC-III timelines of Stage 1 and Stage 2 produced in the cross-comparison above with their method and passed them, unchanged, to our own Stage 3 prompt. Stage 3 is thus entirely ours, applied to their output as an ablation study. The question it asks is not whether their method summarizes well, but whether the timelines it produces can support a summary anchored to dates, which is the goal of our pipeline.

Failure modes. Their rows carry a signed hour offset (e.g. "*event | -24*" instead of "*[*1 day ago from a known event* event*") and nothing else: no date, and no indication of how firmly the timing is known. Our Stage 3 needs both, since it writes a stated date as a date, marks an inferred one as approximate, and places an undated prior event before the admission. With none of these available, the summarizer agent at Stage 3 is forced to infer dates of its own. Two-thirds of the temporal errors occur where the source supported a window of days or weeks, while the summary named a fixed single day inside it. In over half, that date is stated with no hedge at all. A further group stems from the offset of zero, which marks the admission and so cannot separate the admission from the events that caused it, pushing events across that boundary in both directions.

Table S12.2.

Dimension	W&W band	W&W rate	CliniCIRCA band	CliniCIRCA rate
A Faithfulness to source	5.00	99.01%	**5.00**	99.93%
B Temporal accuracy	2.73	62.89%	**4.75**	95.33%

Stage-3 summarization scored by our judge, with our Stage-3 prompt applied unmodified to each system's Stage-2 timeline (Wang and Weiss have no summarization stage of their own). Bands 1–5, averaged over 52 notes per system.

Analysis. Faithfulness holds at the ceiling (5.00, 99.01%), near CliniCIRCA's 5.00 and 99.93%, indicating that the clinical content transfers intact. Temporal accuracy falls to 2.73 (62.89%) against 4.75 (95.33%), and no baseline document reaches the highest band. This indicates that the clinical events present in the summaries are attached to days the source never asserted.

Once again, we acknowledge that this is not a fair comparison, and we do not offer it as one. Wang and Weiss built no summarization stage, and their representation was never intended to carry the distinctions our Stage 3 consumes. The table measures what their output cannot express, not how well their method performs the task it was designed for. This analysis characterizes the compatibility of the baseline's timeline representation with CliniCIRCA's temporal summarization requirements.

References

1. Wang J, Weiss JC. A Large-Language Model Framework for Relative Timeline Extraction from PubMed Case Reports. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2025;2025:598–606.

2. Weiss JC. LLM-Timeline-PMOA: A LLM Framework for Relative Timeline Extraction from PubMed Case Reports [Internet]. 2025. Available from: https://github.com/jcweiss2/LLM-Timeline-PMOA