Additional File 5

Stage 3 Automated and Human-in-the-loop Evaluation Design

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. 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 the following tables.

Table S5.1. Stage-3 evaluation rubric for Dimension A: Faithfulness to Source.

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 S5.2. Stage-3 evaluation rubric for Dimension B: Temporal Accuracy.

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.

Human-in-the-loop 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. 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 the tables in this document.

