Additional File 3

Stage 1 & 2 Automated and Human-in-the-loop Evaluation Design

Evaluation Rubrics

Stage 2's output is evaluated along four dimensions: completeness, semantic faithfulness, temporal-tag accuracy, and anti-hallucination. Completeness assesses coverage of source-grounded clinical events; semantic faithfulness measures preservation of meaning, including negation, uncertainty, quantities, and qualifiers; and temporal-tag accuracy evaluates both the assigned certainty category and resolved date or range. Anti-hallucination identifies unsupported claims and was assessed only by human annotators. Because LLM-as-judge reliability varies across evaluation tasks and remains imperfect for hallucination detection [1, 2], we reserved this criterion for exhaustive human review to avoid contaminating the clinician-verified reference set with missed unsupported claims.

Table S3.1. Stage-2 evaluation rubric for Dimension A: Completeness.

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 S3.2. Stage-2 evaluation rubric for Dimension B: Semantic Faithfulness.

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 S3.3. Stage-2 evaluation rubric for Dimension C: Temporal Tag Accuracy.

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.
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:

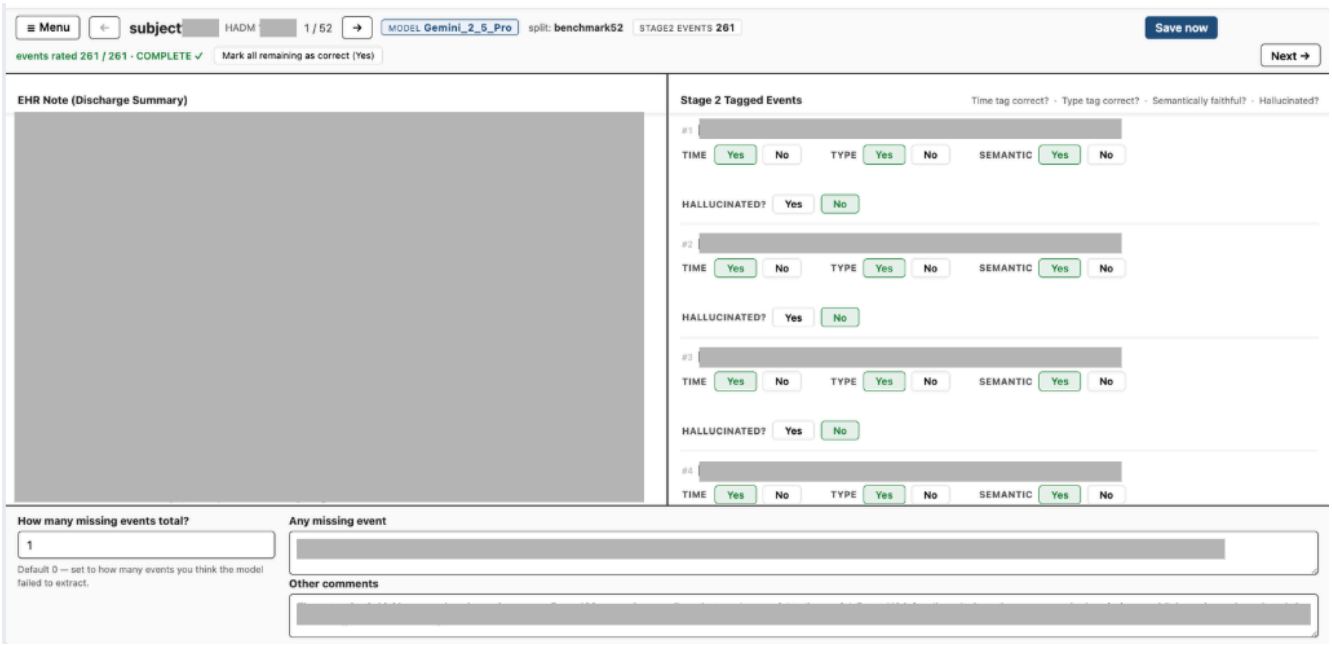
	• tag_over_specific: [EXACT] or [APPROX] was assigned where [PRE ADM] or [INDETERMINATE] was correct, meaning that a date was invented. • tag_under_specific: [PRE ADM] or [INDETERMINATE] was assigned where a date was recoverable. • tag_exact_approx_confused: [EXACT] and [APPROX] were interchanged. • 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 S3-4. Stage-2 evaluation rubric for Dimension D: Anti-Hallucination.

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.

Stage 1 & 2 Human Evaluation Platform

Human evaluation of Stage-2 outputs was conducted through a purpose-built local web application. 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.



Bibliography

1. 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: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vienna, Austria: Association for Computational Linguistics; 2025. p. 238–55. https://doi.org/10.18653/v1/2025.acl-short.20.

2. Li J, Cheng X, Zhao X, Nie J-Y, Wen J-R. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 6449–64. https://doi.org/10.18653/v1/2023.emnlp-main.397.