Additional File 8

Agreement Analysis for Stage 2

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

Table S8.1. Agreement between the Gemini 3.1 Pro and gold annotation.

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

The numbers are calculated over 52 discharge summaries producing 15,891 events. AC1 denotes Gwet's AC1; Pos. and Neg. denote positive and negative agreement on the rare error and dominant non-error classes, respectively. Anti-hallucination is excluded because it was evaluated only by human annotators.

Temporal Error-Type Analysis for Stage 2

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

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

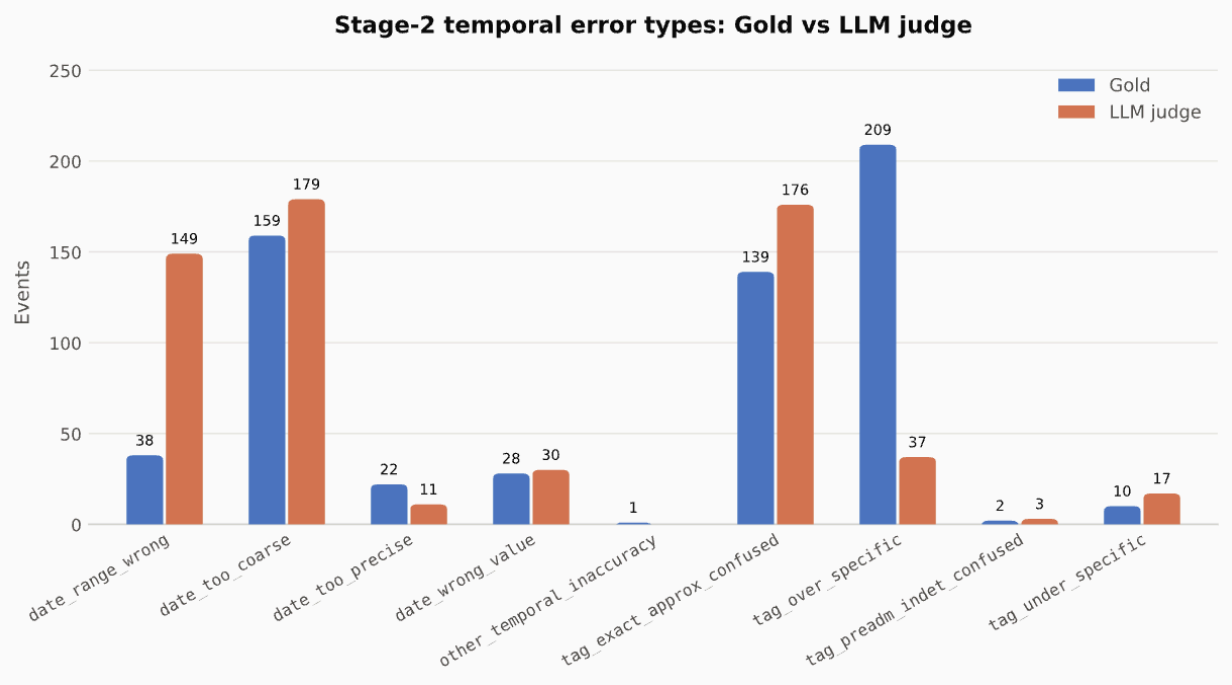


Figure S8.1. Stage 2 Temporal Error Types

The clearest asymmetry concerns over-specified temporal tags. Human annotation identified 175 such errors involving pre-admission medications (151 events) and patient demographics (24 events), of which the judge failed to flag 169 (96.6%). These are not cases in which the human and judge applied different error codes to the same behavior: the judge returned no temporal error at all. The pattern therefore represents a systematic false-negative class in which the judge accepts admission-anchored [EXACT] or [APPROX] labels for standing or pre-admission facts that should receive [PRE ADM].

Other Error Types and Analysis for Stage 2

The 21 non-temporal errors were minor, affecting only 0.13% of 15,891 events. Most reflected over-precision rather than clinical misunderstanding: the model sometimes converted uncertainty into firmer assertions, added unsupported medication details, or misread de-identification masks and OCR-garbled text. The 7 hallucinations similarly involved details not fully grounded in the source, while only one event was missed entirely. Overall, outside temporal tagging, the model was highly faithful, with residual errors arising mainly from resolving ambiguity rather than preserving it.