Additional File 6

Token Usage for Silver-Standard Supervision

To quantify the inference required to construct the silver-standard supervision used in Section 2.6, we recorded the API token-usage metadata returned by Gemini 2.5 Pro for the 1,000 admissions processed through the frozen three-stage framework. Across Stages 1 to 3, generation processed 45.13M input tokens and produced 16.70M output tokens, for a combined total of 61.84M tokens, or 61,836 tokens per admission on average. Stage 1 accounted for 25.66M tokens (41.5% of the total), Stage 2 for 19.52M (31.6%), and Stage 3 for 16.66M (26.9%).

Table S6.1. Token usage for generating the 1,000-example silver-standard corpus with the frozen Gemini 2.5 Pro.

Stage	Input	Output	Total
Stage 1: Atomic Clinical Event Extraction	20,464,371	5,198,803	25,663,174
Stage 2: Date & Certainty Tagging	9,794,222	9,723,468	19,517,690
Stage 3: Temporally Guided Summarization	14,874,581	1,780,523	16,655,104
Total	45,133,174	16,702,794	61,835,968

Total tokens are the sum of input and output tokens. The grand total corresponds to an average of 61,836 tokens per admission across all three stages. Of the input tokens, 16,177,254 were served from cached context and are already included in the input counts.

These counts quantify the inference used to generate the 1,000-example silver corpus in Section 2.6 and illustrate the computational demands of processing complete discharge summaries and their intermediate representations through a multi-stage long-context pipeline. The resulting corpus, however, can be reused across three downstream tasks and multiple open-weight student models, amortizing the initial teacher-generation cost. Of the reported input tokens, 16,177,254 were served from cached context; because cached tokens are a subset of input tokens, they are not added again to the total.