Additional File 1

Pipeline Agent Candidate Selection

Following the selection of agents in Section 2.3.1, a candidate is eligible if it supports a native context window of at least 32K tokens which is sufficient to fit our P99 input without truncation. All candidates are evaluated with identical prompts and greedy decoding ($t = 0$).

An open-weight candidate additionally has to be (i) downloadable under a license permitting research and on-premise inference and (ii) loadable on our deployment stack across four L40S GPUs at bfloat16 precision. A proprietary candidate additionally has to be (iii) served through a managed endpoint compatible with our institutional data-use agreement (e.g., Amazon Bedrock or Google Vertex AI) and (iv) cost-feasible at corpus scale. A complete table of agent specifications can be found in the following table.

Table S1.1. Pipeline Agent Candidates Information.

Agent Model	Provider	Params (total/active)	Arch.
Gemini 2.5 Pro [1]	Google (via Vertex AI)	Undisclosed	n/r
Gemini 2.5 Flash [1]	Google (via Vertex AI)	Undisclosed	n/r
Claude Sonnet 4.6 [2]	Anthropic (via Bedrock)	Undisclosed	n/r
Claude Haiku 4.5 [3]	Anthropic (via Bedrock)	Undisclosed	n/r
Llama 3.3 70B [4]	Meta	70B	Dense
Qwen 3.6 35B-A3B [5]	Alibaba	35B / 3B	MoE
Gemma 4 26B-A4B) [6]	Google	26B / 4B	MoE
MedGemma 27B [7]	Google	27B	Dense
Mistral Small 3.2 24B [8]	Mistral AI	24B	Dense
Meditron 3 (Qwen 2.5 7B) [9]	OpenMeditron/EPFL	7B	Dense
Phi-4 Mini 3.8B [10]	Microsoft	3.8B	Dense

Benchmark Size Selection Via Power Analysis

We fix the number of manually evaluated documents using an a priori power analysis for a chi-square test. The required sample size is

$$N = \frac{\lambda^*}{\omega^2},$$

where ω is Cohen's effect-size index and λ^* is the noncentrality parameter at which the noncentral chi-square distribution $\chi^2(df, \lambda)$ reaches the target power relative to the critical value $\chi^2_{1-\alpha}(df, \lambda)$. For a pairwise comparison with one degree of freedom, we targeted a large effect size ($\omega = 0.5$), a significance level of $\alpha = 0.05$, and power of $1 - \beta = 0.95$. This gives

$$\lambda^* = (z_{1-\alpha/2} + z_{1-\beta})^2 = 12.99,$$

and therefore

$$N = \frac{12.99}{0.5^2} = \frac{12.99}{0.25} = 51.96$$

We round this value up to $N = 52$ documents. Applying a finite-population correction for the full corpus of 14,882 documents produced an adjusted estimate of $N = 51.96$, leaving the required sample size effectively unchanged. At $N = 52$, the achieved power is 0.950.

The analysis is conducted using the *pwr* package and independently verified using the noncentral chi-square distribution implemented in *SciPy*. The sampling unit is the discharge-summary document. Thus, the independence assumption applies across sampled documents, and correlations among events within a document do not artificially increase the effective sample size.

Reference

1. Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. 2025. https://doi.org/10.48550/arXiv.2507.06261.

2. Anthropic. Claude Sonnet 4.6 System Card. 2026.

3. Anthropic. Claude Haiku 4.5 System Card. 2025.

4. Meta AI. Llama 3.3 70B Instruct Model Card. 2024.

5. Qwen Team. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All. 2026.

6. Gemma Team. Gemma 4 Technical Report. 2026.

7. Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al. MedGemma Technical Report. 2026.

8. Mistral AI. Mistral-Small-3.2-24B-Instruct-2506 Model Card. 2025.

9. OpenMeditron. Meditron3-Qwen2.5-7B Model Card. 2025.

10. Abouelenin A, others. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. ArXiv Prepr ArXiv250301743. 2025.