

Reproducibility Statement

Our paper “Towards Automatic Evaluation and Selection of PHI De-identification Models via Multi-Agent Collaboration” ensures reproducibility of its AI-generated results by fully specifying the proposed TEAM-PHI framework, including formal definitions of the multi-agent evaluation function, prompt templates, and strict JSON output formats for De-id models, Evaluation Agents, and LLM majority vote, which standardize inputs and outputs for automated scoring and majority voting. All implementation details and prompts are released in an anonymous public repository (<https://anonymous.4open.science/r/SAFE-0915>), and Appendix B.2 provides direct URLs to every model used (e.g., Llama-3-8B/70B, GPT-3.5, GPT-4o, Mistral-7B, Gemma-2, LPPA4k/5k). In this work, we use 100 fully annotated clinical notes provided by a major U.S. hospital, in which all PHI entities were manually identified by medical experts. These notes reflect real-world clinical documentation and include a wide range of patient information, and more details are provided in Appendix B.1. Due to the sensitive nature of medical data and privacy regulations, these clinical notes are not publicly available. For questions regarding data use or evaluation protocols, please contact the author team. The paper defines all evaluation metrics—precision, coverage, correct count, and recall-proxy—in Appendix B.3, and ground-truth and human evaluations in Appendix B.4 confirm that the automated rankings align with supervised and expert judgments. Appendix B.5 discloses the compute resources, including dual NVIDIA H100 GPUs (80 GB), Apple M2 Pro hardware, Azure HIPAA-compliant inference for GPT models, per-run and total compute time (under two hours per run; about 200 GPU-hours total), ensuring others can match experimental conditions.

By coupling transparent methodological and computational reporting with discussion of limitations such as dataset size and prompt sensitivity (Appendix C), and by providing the complete code and prompts in an anonymous public repository (<https://anonymous.4open.science/r/SAFE-0915>), our work offers all essential information required for independent reproduction of its results.