

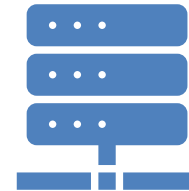
Evaluating Local LLM-Based De-identification in a Hybrid Clinical RAG Framework



Arun Joshi, Aayush Adhikari
University of Texas at Austin
https://github.com/aayushadhikari/fargo_synthea

Abstract

- We present a hybrid architecture that uses local LLMs for PHI removal before sending data to remote LLMs for clinical reasoning.
- Two approaches compared: Fine-tuned Gemma 2B (LoRA) vs. Baseline Gemma 3B.
- Findings: Local LLMs can help but fail to achieve 100% PHI removal, requiring more robust solutions.



Introduction

Problem: HIPAA regulations limit use of powerful LLMs in healthcare.

Goal: Combine privacy of local models with reasoning power of remote models.

Approach: Local LLM removes PHI →
Remote GPT-4.1 provides clinical reasoning.

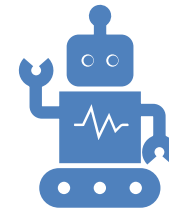


System Architecture

Remote Orchestrator (GPT-4.1)

Local De-identification Models: Fine-tuned
Gemma 2B, Baseline Gemma 3B

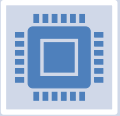
Clinical Data Tools (8 retrieval functions)
Strict separation of PHI handling from
remote reasoning.



Fine-Tuning Process



Dataset: 343 synthetic clinical examples.



Generated using Faker + GPT-OSS:20B for realistic contexts.



LoRA Config: rank=8, alpha=16, 3 epochs, batch size 2.



Goal: Improve PHI recall while preserving clinical meaning.

Evaluation Methodology



Manual inspection of outputs for sample queries.



Validated tool-calling workflow and orchestration logic.



No formal quantitative metrics yet.



Key metric: PHI removal completeness vs. clinical fidelity.

Results – Fine-tuned Gemma 2B

- High recall for PHI removal.
- Over-redaction of non-PHI content.
- Occasionally removed useful clinical details.
- Better pattern recognition than baseline.



Results – Baseline Gemma 3B

- Preserved more clinical details.
- Lower PHI recall; missed identifiers.
- Inconsistent redaction.
- Higher compliance risk due to misses.



Clinical Analysis Example

- Query: Pregnancy-related medication safety.
- Workflow: Correct tool calls → Accurate medication guidance.
- Limitation: Upstream PHI removal removed relevant allergy info.
- Outcome: Safe but incomplete advice.



Conclusion

Hybrid architecture is viable for HIPAA-compliant AI.

Current LLM-only de-identification is unreliable for production.

Need hybrid de-identification (rule-based + NER + LLM) for robustness.

Future Work



Hybrid de-identification pipeline.



Evaluate advanced open-source models.



EHR integration via FHIR.



Formal clinical validation with clinicians.



Human-in-the-loop privacy auditing.