Gemedi

Discharge Medical Note Generator

Zejun Zhou, Yongcheng Shi, Yumian Cui, Ruoqian Zhang

CSCI 1470

Introduction

The Problem: The Data Gap

- **HIPAA Restrictions:** Real medical notes containing PHI are inaccessible for research.
- **Sanitization Loss:** Public datasets (MIMIC-IV) lack the authentic **PHI structure** and **flow** needed for training realistic clinical AI.

Motivation: Limitations of Standard LLMs

- Our baseline tests with **Llama 3.1 8B** revealed a critical failure mode:

Setup	Format Realism	Medical Reasoning
Standard Prompt	✅ Structured	❌ Poor
Expert Prompt	❌ Hallucinated	✅ Reasonable

Our Approach We propose a **Dual-Discriminator Feedback Loop**:

1. **Generate:** Create synthetic notes using a fine-tuned Generator.
2. **Critique:** Evaluate using separate **Realism** (Format) and **Reasoning** (Logic) discriminators.
3. **Refine:** Update generation via iterative prompt feedback.

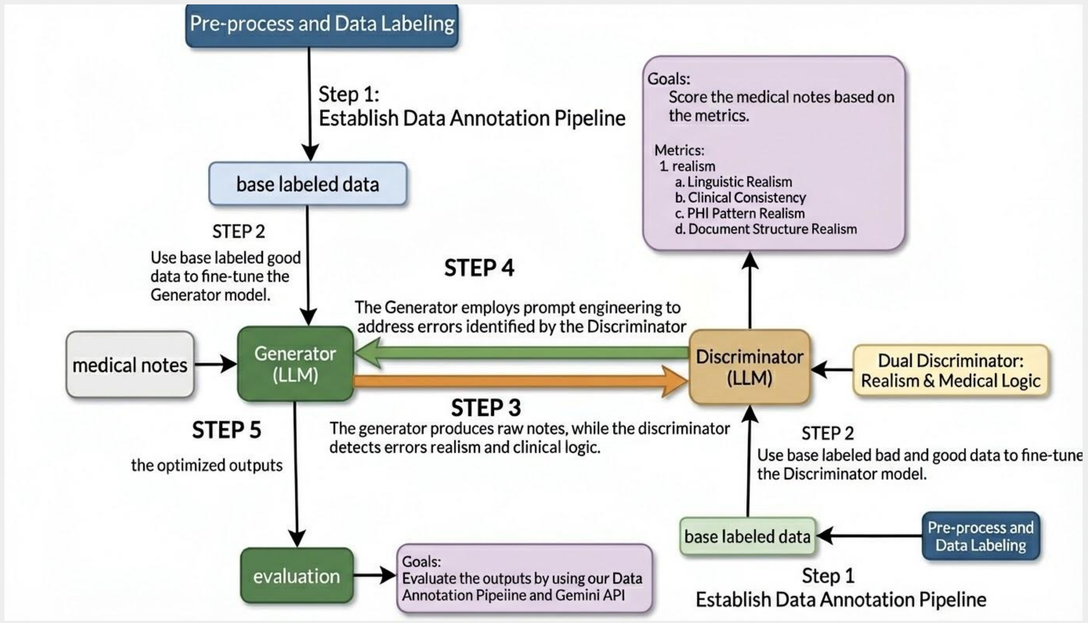
Data

Source: MIMIC-IV-2 (De-identified).

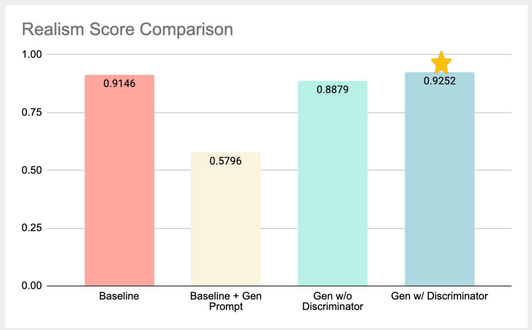
Synthetic Entity Injection Strategy: We repopulate de-identified slots with **synthetic PHI** to create two variations:

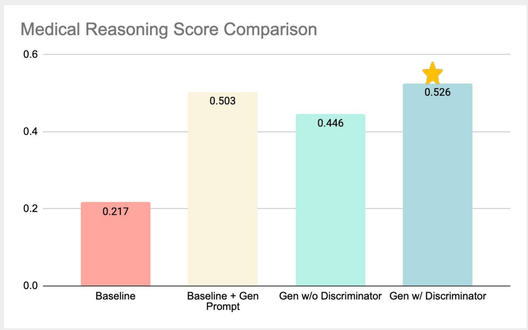
- **Good Data (Consistent):**
  - Injected with **context-aware, consistent PHI** to simulate realistic, error-free notes.
  - **Target:** Generator Fine-tuning & Realism Discriminator Fine-tuning.
- **Bad Data (Flawed):**
  - Injected with **logical errors** (e.g., further date)
  - Includes **Reasoning Labels** explaining the specific flaws.
  - **Target:** Realism Discriminator Fine-tuning.

Architecture



Results





Gemedi: **significantly improved** realism, **comparable-to-superior** medical reasoning.

References

Self-Refine – Madaan et al., 2023. *Self-Refine: Iterative Refinement with Self-Feedback.* arXiv:2303.17651.

QLoRA – Dettmers et al., 2023. *QLoRA: Efficient Finetuning of Quantized LLMs.* arXiv:2305.14314.

Self-Instruct – Wang et al., 2022. *Self-Instruct: Aligning Language Models with Self-Generated Instructions.* arXiv:2212.10560.

Challenges

Resource & Memory Bottlenecks

- **Problem:** Performing full-parameter fine-tuning on Llama-3.1 was computationally prohibitive. It required massive VRAM that exceeded our **single-GPU** setup and resulted in extremely slow convergence.
- **Solution:** We implemented **QLoRA** (Quantized Low-Rank Adapters). This technique allowed us to **freeze the massive pre-trained base model** and only update small, trainable adapter layers. This reduced memory usage by over 90%, enabling efficient training on limited hardware without sacrificing performance.

The "Reasoning" Gap

- **Problem:** Our initial fine-tuned discriminator was limited to detecting **surface-level linguistic and formatting issues** (e.g., invalid discharge medical notes formats). It lacked the deep clinical knowledge required to catch **semantic medical contradictions**, such as a diagnosis of "Hypertension" followed by a prescription for "Insulin."
- **Solution:** To bridge this gap, we benchmarked multiple LLMs and integrated the **Google Gemini API** as a dedicated second evaluator. It demonstrated superior performance in identifying complex logical flaws and providing actionable clinical feedback that our smaller local model missed.

Future Work

- **Update Model Weights directly**: Instead of just changing the prompt, we want to use the Discriminator feedback to train the Generator. This makes the model itself smarter, not just the output.
- **Build Our Own Medical Judge**: Train a local model to replace the **Google Gemini API**. This will save money (no API cost) and run much faster.
- **Use Bigger, Smarter Models:** With more powerful GPUs, we plan to fine-tune larger models (like **Llama-3 70B**) to get even better medical reasoning than our current 8B model.
- **Verify Real-World Utility:** To prove our data is actually useful, we will train external NLP models using only our synthetic notes. If these models perform well on real data, it proves our synthetic data can replace real sensitive data.