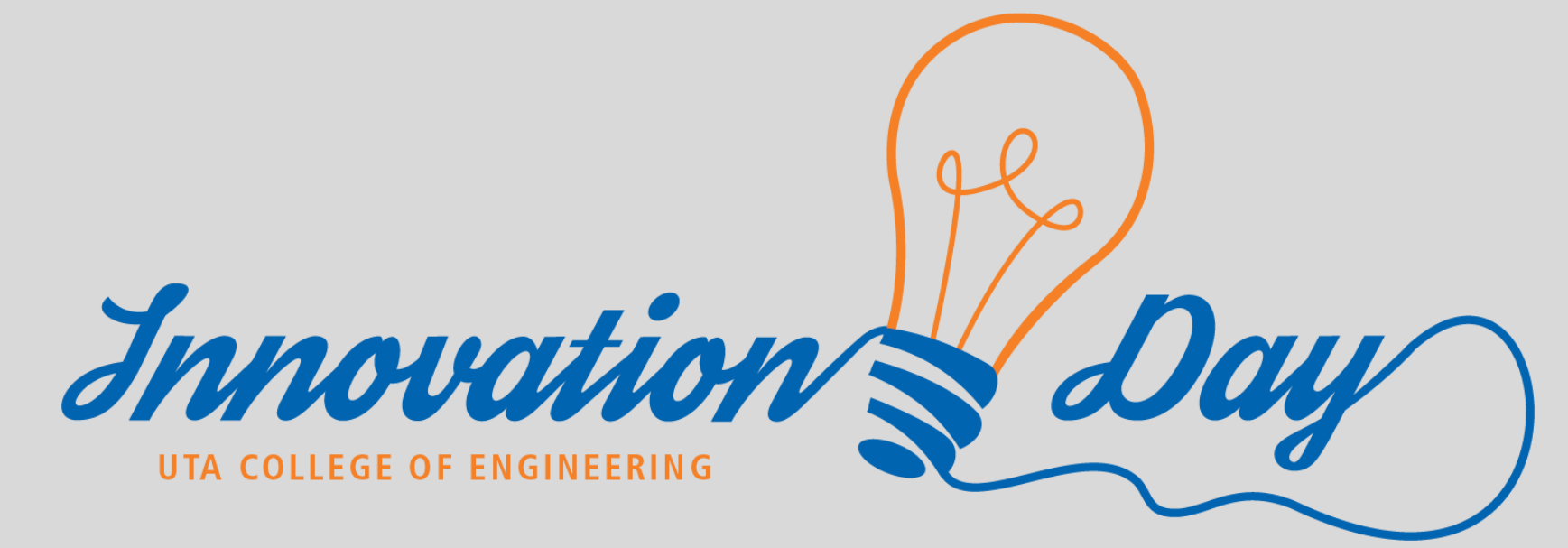


Probing AI Safety: How LLMs Handle Sensitive Medical Prompts and Misinformation

Abigia Abate | Mikaela Maramba | Faculty: Faysal Hossain Shezan



Executive Summary

Large language models (LLMs) are increasingly used for medical question answering, raising concerns about privacy, safety, and misinformation. This project evaluates how LLMs handle sensitive medical prompts containing synthetic patient data. We developed an automated pipeline that generates diverse medical questions, produces model responses, and evaluates outputs using a MIRIAD-based reference system with cosine similarity and BLEU metrics. The system also explores privacy-preserving transformations (Papillon) to measure tradeoffs between data protection and response quality. Results highlight challenges in evaluating medical correctness, particularly when traditional metrics fail to capture semantic alignment, and demonstrate the need for scalable safety evaluation frameworks.

Background

- Large language models are increasingly used for healthcare-related questions, but they are not regulated medical systems and may generate unsafe or misleading advice. Additionally, these models may mishandle sensitive personal data, raising concerns related to privacy regulations such as HIPAA. Existing evaluation approaches are limited in scale and often fail to capture real-world variability in medical queries. This project addresses the need for automated, scalable evaluation of LLM behavior in sensitive medical contexts.

Experimental Setup

We developed an automated evaluation pipeline for testing LLM behavior on medical questions. A disease–symptom dataset from Kaggle is used to generate diverse medical prompts with variations in tone, structure, and context. These prompts are passed to LLMs (Gemini and LLaMA via Ollama on a GPU server) to generate responses. Each response is evaluated using a MIRIAD-based retrieval system, which identifies the most semantically similar medical reference and compares outputs using cosine similarity and BLEU scores. Results are stored in structured tables for further analysis. The system is designed to support scalable batch processing and future integration of privacy-preserving transformations (Papillon).

Table 1. Table for fictional profile and prompt injection setup

Profile attribute	Ex
Age(range)	45
Condition(list conditions)	Type II diabetes
Medication	insulin
Name(identifiers)	Name, DOB
Prompt tone	Formal/casual
Level of detail	Specific/vague

Experimental Test Plan

- We evaluated LLM responses across multiple conditions using both manually generated and automated prompts.
 - 50-100 automated prompts generated from disease–symptom data
 - Additional manually curated prompts for validation
 - Models tested: Gemini (API-based) and LLaMA (via Ollama, in progress)
 - Metrics: cosine similarity and BLEU score against MIRIAD references
- The goal was to assess response quality, consistency, and alignment with medical knowledge. Future experiments will include Papillon-based privacy transformations and larger-scale batch testing.

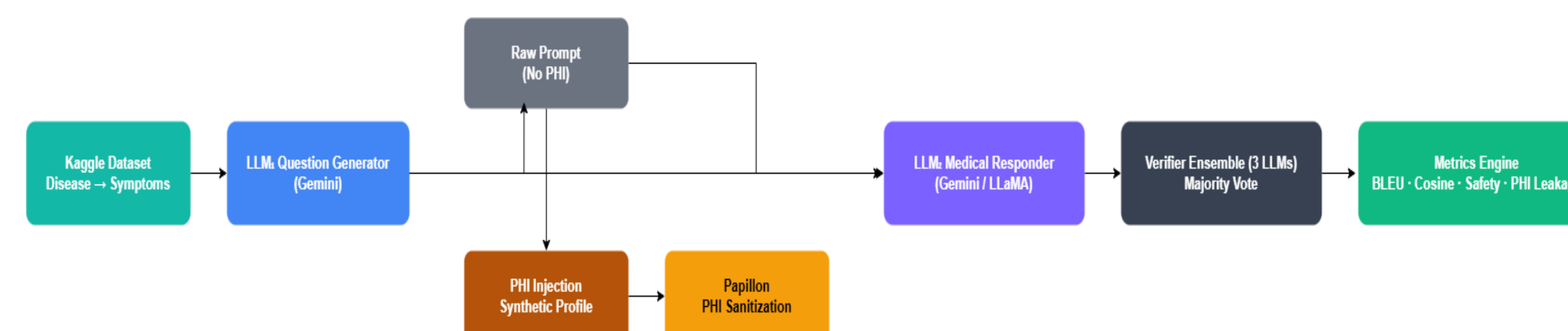


Figure 1. Automated evaluation pipeline: disease-symptom prompts are passed to LLMs with optional PHI injection, then scored against the MIRIAD corpus using cosine similarity and BLEU.

Experimental Results

- LLM responses demonstrated strong semantic alignment with medical references, reflected by high cosine similarity scores (mean: 0.88, range: 0.79–0.98). BLEU scores frequently failed to capture this alignment—often return ingnear-zero values despite clinically reasonable outputs (mean BLEU: 0.31, median:0.13). This reveals a critical limitation of surface-level evaluation metrics for medical AI.
- The automated pipeline processed 50–100 prompts efficiently, confirming scalability improvements over manual evaluation. These findings support cosine similarity as the preferred evaluation metric for medical LLM correctness.
- Papillon effectively redacted PHI identifiers (names, SSNs, locations). In some cases sanitized prompts produced equally correct responses; in others, removal of clinical context reduced relevance—revealing a measurable privacy–accuracy trade off.
- Manual evaluation revealed inconsistent PHI handling across models: GPT-4 and Claude declined to repeat sensitive identifiers in most cases, while Gemini occasionally reproduced SSN digits when asked to summarize prior conversation context.

Conclusions

- This work demonstrates the feasibility of a scalable, automated framework for evaluating LLMs in sensitive medical domains. Results reveal a critical gap in current evaluation methods, where widely used metrics fail to reflect true semantic correctness.
- The proposed pipeline provides a foundation for large-scale testing of LLM safety, correctness, and privacy behavior. Future work will integrate Papillon for privacy analysis and expand multi-model comparisons, moving toward more robust and reliable medical AI evaluation systems.

References

- [1] Lehman, E., et al. "MIRIAD: Medical Information Retrieval and Inference Assessment Dataset." (2023). [2] Dhivyes, R. "Diseases and Symptoms Dataset." Kaggle (2024). [3] Zhou, L., et al. "Papillon: PHI Redaction Framework for LLMs." arXiv (2024). [4] U.S. Dept. of HHS. "HIPAA Privacy Rule." 45 CFR Parts 160 & 164.