Information Privacy Controls for Corporate Use of Public LLMs

Data Science Institute
COLUMBIA UNIVERSITY

Jiawen Song, Kaiyuan Liu, Longcong Xu, Shaonan Wang, Shukai Wang

Mentor: Liang Tang

Data Science Capstone Project
with General Electric

Introduction

This project aims to establish robust information privacy controls in corporate environments when using Large Language Models (LLMs). We address the concern of exposing private data to public LLMs by focusing on de-identification and anonymization processes, ensuring compliance with privacy regulations. Our approach involves anonymizing Protected Health Information (PHI) from Electronic Health Records (EHRs), demonstrating the LLM's adaptability in diverse privacy-sensitive contexts, including aviation safety data.

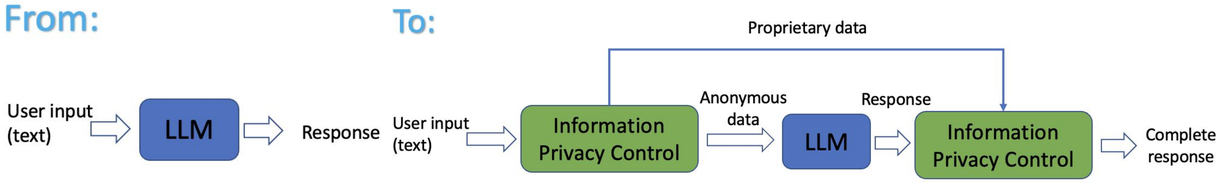

Figure 1. Workflow Shift of Information Privacy Control Mechanism

Methodology

Datasets

The data used for de-identifying medical records is n2c2 NLP Research Data Sets (i2b2) with 1300+ free-text records of patient diagnoses, treatments, etc. Each record has both medical annotation text and PHI tags. The Aviation Safety Reporting System (ASRS) database used for transfer learning contains text narratives of aviation incidents with details like airport and aircraft names sanitized.

Models

LSTM (Local Deep Learning Model)	• Implemented in Python with Keras library on Google Colab • Constructed a bidirectional LSTM to classify whether each word belongs to PHI tags • Applied word embedding to transform to vector representations to feed in the model
GPT (Reference Model)	• Implemented on Google Colab via OpenAI API • Established GPT-4 model with both implicit and explicit instructions as system prompts • Feed in the i2b2 clinical text via normal prompt and performed model evaluation • Also fine-tuned GPT 3.5 specifically based on i2b2 dataset and re-tested the results
Llama2 (Run Locally)	• Quantized Llama2 13B: Tested the model via Google Colab, but found unexpected changes in original text and incomplete generations of results • Llama2 70B-Chat: Tested the model via HuggingChat. We set the explicit prompt as system prompt and input clinical text via chat directly to derive de-identified results

Prompt Engineering for Llama 2 and GPT

Implicit Prompt	Explicit Prompt
Replace private information such as a person's name, date, birthday, gender, location, etc. with #####.	Please anonymize the following clinical note. Replace all the following information with the term "########": Redact any strings that might be a name or acronym or initials, patients' names, doctors' names, the names of the M.D. or Dr., redact any pager names, medical staff names, redact any strings that might be a location or address, such as "3970 Longview Drive", redact any strings that look like "something years old" or "age 37", redact any dates (like 2081-9-29) and IDs and record dates, redact clinic and hospital names, redact professions such as "manager", redact any contact information. Don't cover up some of the medical descriptions like blood pressure.

Results

Our evaluation using i2b2 medical records revealed that GPT-4 with explicit prompts outperformed the one with implicit prompts across all metrics, notably in recall, which is essential for complete data anonymization. Fine-tuning GPT-3.5 further enhanced its performance, achieving near-perfect scores, indicating its superior effectiveness in identifying and anonymizing private information.

Llama2 70B-Chat demonstrated better performance in de-identifying private information with implicit prompts compared to explicit prompts. However, both settings fell short of the performance achieved by GPT models. In the context of aviation data, Llama2 showcased acceptable accuracy and recall but lacked precision. LSTM was implemented to compare LLMs with traditional deep learning methods.

Method	Prompt	Accuracy	Recall	Precision	F-1 Score
LSTM	—	91.01%	77.73%	63.99%	70.19%
GPT-4	Implicit	96.44%	75.94%	73.63%	74.77%
GPT-4	Explicit	97.44%	95.03%	74.81%	83.72%
GPT-3.5 fine-tuned	Explicit	99.97%	99.27%	100.00%	99.61%
Llama2 70B-Chat	Implicit	89.48%	59.20%	46.00%	45.80%
Llama2 70B-Chat	Explicit	85.93%	58.01%	27.08%	36.92%
Llama2 70B-Chat (Transfer Learning)	Explicit	96.47%	72.53%	12.38%	21.15%

Figure 2. Model Performance on Medical Record De-identification and Transfer Learning on Aviation Data

Conclusion

In addition to traditional deep learning models, LLMs can be employed for de-identifying private information in text records, provided that LLMs can run locally or in a protected environment. Prompt engineering for LLMs often involves a trial-and-error process, requiring precise composition for specific tasks. Due to limited computing resources available for this project, Llama 2's performance falls short in de-identification tasks compared to GPT-4 and fine-tuned GPT-3.5 models. Future work on Llama2 could focus on enhancing system prompts and incorporating fine-tuning techniques.

Acknowledgments

We express our sincere gratitude to our mentor, Liang Tang, at General Electric and the Data Science Institute at Columbia University for their invaluable collaboration and support throughout this Capstone project.

References

Ahmed, T., Aziz, M. M. A., & Mohammed, N. (2020). De-identification of electronic health record using neural network. *Scientific reports*, *10*(1), 18600.

Aviation Safety Reporting System. (n.d.). ASRS Database Online [Data set]. Aviation Safety Reporting System (NASA). https://asrs.arc.nasa.gov/