Information Privacy Controls for Corporate Use of Public LLMs

by

Jiawen Song (js6123)

Kaiyuan Liu (kl3447)

Longcong Xu (lx2305)

Shaonan Wang (sw3720)

Shukai Wang (sw3715)

Final Report

Working with GE Aviation and
Data Science Institute at Columbia University

December 16, 2023

Table of Contents

Contents Page #

1. Introduction

This Capstone project represents a collaborative effort between the Data Science Institute at Columbia University and General Electric (GE). The central aim of this project is to develop an Information Privacy Control mechanism in response to the concerns of exposing private data while using public online Large Language Models (LLM) in corporate settings. This initiative addresses this critical problem of how to maintain data privacy and security while applying public LLM in compliance with corporate policies by anonymizing private information from input data. Given the existing constraints of conventional privacy techniques and the distinctive risks of using public LLM within corporations, we were motivated to use and finetune GPT and Llama2 models to aid corporations in anonymizing data before applying public LLMs for the protection and assurance of data privacy.

Building upon the foundational contributions of previous research, our project draws inspiration from these pioneering efforts. Dernoncourt et al. introduced a groundbreaking de-identification system employing artificial neural networks (ANNs), which uniquely operates without the need for manually crafted features or rules.[1] Similarly, the innovative approach of Yadav, Shweta, et al., with their DI-RNN system, leverages deep learning for patient data de-identification (PDI).[2] Additionally, the work of Tang et al., which involves generating high-quality synthetic data with ChatGPT and fine-tuning a local offline model for clinical text mining, has also been influential.[3] Our work extends these concepts by applying and comparing the Llama2 and GPT models, furthering the advancements in the field and contributing to the evolution of de-identification and privacy-preserving techniques in data science.

In this project, the primary focus of our team centers on the utilization of the Llama2 and GPT models within the framework of the 2014 i2b2/UTHealth de-identification challenge dataset. Our aim is to develop suitable prompts that effectively remove protected health information (PHI),

[1] Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc. 2017 May 1;24(3):596-606. doi: 10.1093/jamia/ocw156. PMID: 28040687; PMCID: PMC7787254.

[2] Yadav, Shweta, et al. "Deep learning architecture for patient data de-identification in clinical records." *Proceedings of the clinical natural language processing workshop (ClinicalNLP)*. 2016.

[3] Tang, Ruixiang & Han, Xiaotian & Jiang, Xiaoqian & Hu, Xia. (2023). Does Synthetic Data Generation of LLMs Help Clinical Text Mining?.

adhering to the standards set by the Health Insurance Portability and Accountability Act (HIPAA). The subsequent phases involve the refinement and optimization of GPT to enhance its efficacy. Furthermore, we have engaged in transfer learning, applying the Llama2 model to the Aviation Safety Reporting System (ASRS) dataset after manually restoring the masked information, which encompasses aviation safety narratives.

Our comprehensive comparative analysis of the performance of Llama2 and GPT on both datasets has yielded significant insights, with the fine-tuned ChatGPT-3.5 emerging as the superior performer. This finding underscores the impactful role of both the prompts and the model itself in determining performance outcomes. These elements warrant further exploration in future research, highlighting the meaningful contributions of our work to ongoing advancements in the field.

2. Methods

2.1 Background

The widespread use of electronic health records (EHRs) has produced a substantial volume of unstructured patient medical data accessible, presenting a valuable opportunity for advancing medical research and enhancing healthcare provision. However, according to HIPAA, all the identifying information in patients' medical records needs to be removed in order to protect patient privacy. Therefore, information privacy control method, such as de-identification techniques that completely remove any personal patient details like names, dates, addresses, etc., is highly desirable. This information privacy control method can mask sensitive information in the original text and generate anonymous data before proceeding with further research or feeding into ML or LLM models, which effectively safeguards data privacy and prevents privacy leakage (Figure 1).

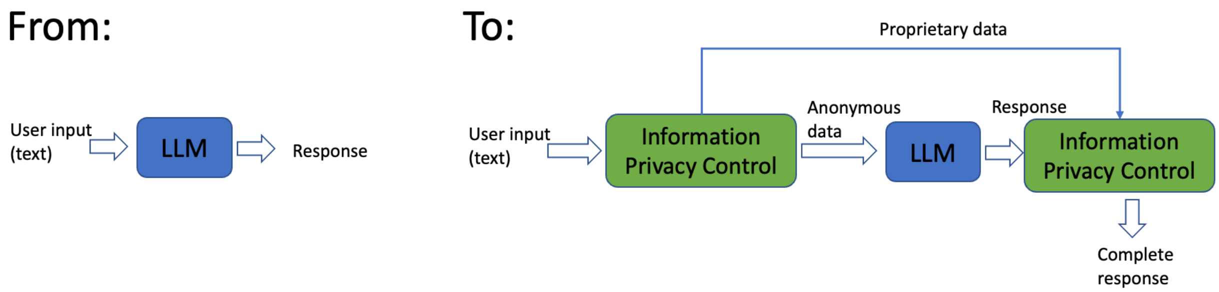


Figure 1: Definition of Information De-identification

Given that most EHR data are unstructured free-form text data, manual de-identification, where people manually delete private information from text, is a time-consuming and error-prone process. As a result, researchers have been actively seeking automatic de-identification methods. Previous studies have introduced a variety of methods for de-identification, including rule-based matching and machine learning-based recognition. Rule-based methods primarily rely on finding pre-defined word patterns using regular expressions and looking up searching dictionaries. Machine Learning-based techniques encompass building machine learning and deep learning models for de-identification and setting up scalable pipelines for large-scale data[4]. With the advent of LLMs in 2023, there is a growing interest in exploring LLM-based de-identification methods and utilizing these models for anonymizing clinical records[5]. Among all the popular LLMs, GPT and Llama2 are two most renowned LLM models developed by Open AI and Meta, respectively. Our primary objective for the first checkpoint is to perform the de-identification task using GPT and Llama2, which could pave the way for subsequent fine-tuning.

[4] Trienes, J., Trieschnigg, D., Seifert, C., & Hiemstra, D. (2020). Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records. arXiv preprint arXiv:2001.05714.

[5] Liu, Zhengliang & Yu, Xiaowei & Zhang, Lu & Wu, Zihao & Cao, Chao & Dai, Haixing & Zhao, Lin & Liu, Wei & Li, Quanzheng & Liu, Tianming & Zhu, Dajiang & Li, Xiang. (2023). DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4.

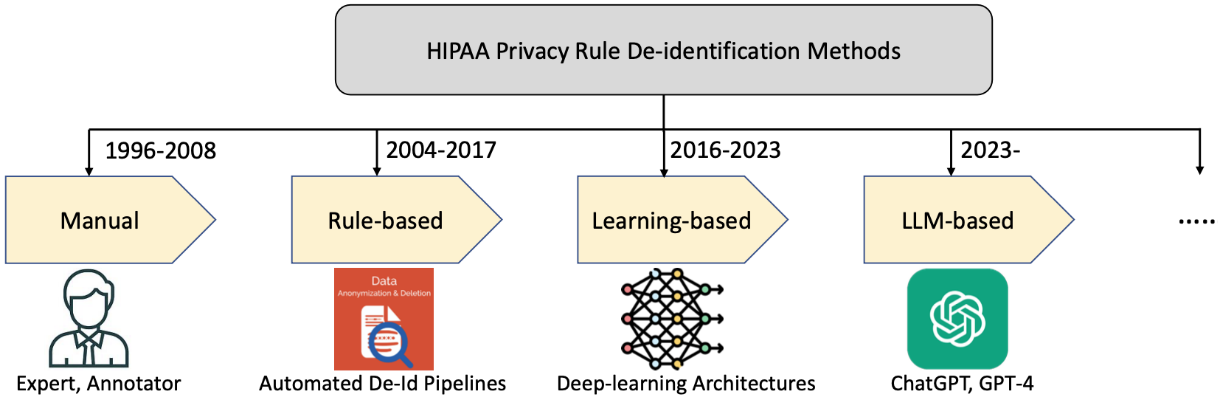


Figure 2: History of De-identification Methods in Accordance with HIPPA

2.2 Data

2.2.1 Dataset Overview

Two datasets, n2c2 NLP Research Data Sets and ASRS database, were used in this project. The performance of GPT and LLama2 models with prompt engineering was tested on the n2c2 testing set, and the n2c2 training set was incorporated during the GPT3.5 finetuning. The experiment of transfer learning was applied to the safety narratives in the ASRS database.

The n2c2 NLP Research Data Sets, which was previously known as i2b2. This clinical data includes a wide range of healthcare-related text sources. These sources may include electronic health records (EHRs), clinical notes, radiology reports, pathology reports, and more. Clinical data is often unstructured and may contain free-text descriptions of patient encounters, diagnoses, treatments, and medical history. The dataset comes from the challenge of the 2014 i2b2 de-identification track, aiming to remove protected health information (PHI) in the medical record. It was later used for various challenges or shared tasks focused on natural language processing and clinical informatics. The n2c2 community organizes these challenges and has addressed various tasks related to clinical text processing. Users can access the data independently through the DBMI Data Portal[6].

[6] "Data Sets." *National NLP Clinical Challenges (n2c2)*, https://n2c2.dbmi.hms.harvard.edu/data-sets. Accessed 27 October 2023.

The n2c2 dataset contains a set of over 1300 patient records, with each record in the format of XML. The training gold set contains 790 records with PHI identified, and the testing set contains 514 records with PHI identified. The same testing records without PHI identified are also available. Each XML file has a root-level XML node <deIdi2b2> which will contain a <TEXT> node that holds the medical annotation text and a <TAGS> node containing annotations, the PHI, for the document text. The specific annotations contained in each file are described by the accompanying DTD file and annotation guidelines.

The Aviation Safety Reporting System (ASRS) database, which will be used for the transfer learning part, contains the text narratives submitted by reporters, with identifying details, like aircraft registration codes and airport codes, sanitized by notations like XXX and “Aircraft Y”. We downloaded approximately 260,000 records from January 2019 to September 2023 in the CSV format. Each record contains information including date, place, environment, aircraft, person, event assessment, narrative, or synopsis (text). For this project, we are only interested in the narrative part.[7]

2.2.2 Exploratory Data Analysis

The testing set in the n2c2 dataset was used for model experiments and comparisons. The XML files in this dataset consist of two parts, the raw text of medical records and tags containing the PHI categories and content of the privacy information being anonymized. We look into these two parts with visualizations to understand the dataset.

[7] ASRS Database Online - Aviation Safety Reporting System.” *NASA ASRS*, https://asrs.arc.nasa.gov/search/database.html. Accessed 27 October 2023.

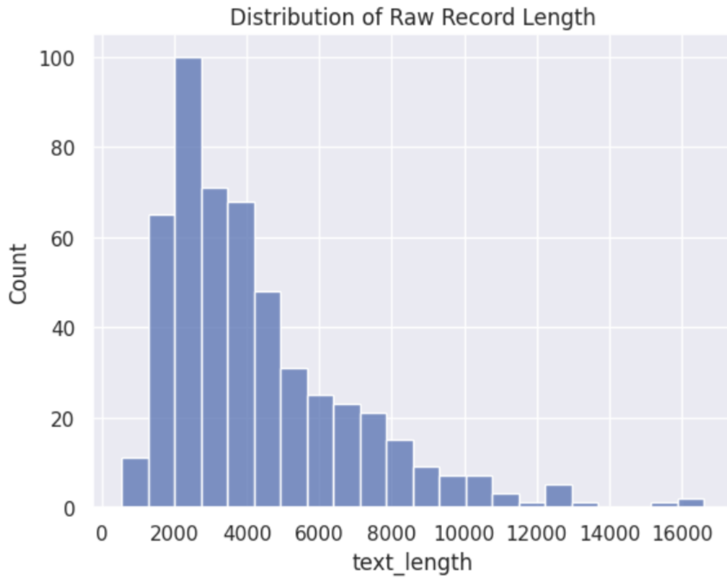


Figure 3: Histogram of Medical Record Lengths

The histogram of medical record lengths is distributed as skewed to the right in Figure 3. Most medical records are of length between 2000 to 6000 words, with a long tail between 6000 and 16000 words. Since the LLM models we use are prompt-based, this indicates that records with extremely longer lengths than the majority may slow down the overall model running process, or reach the limit of max tokens that can be input in the prompt.

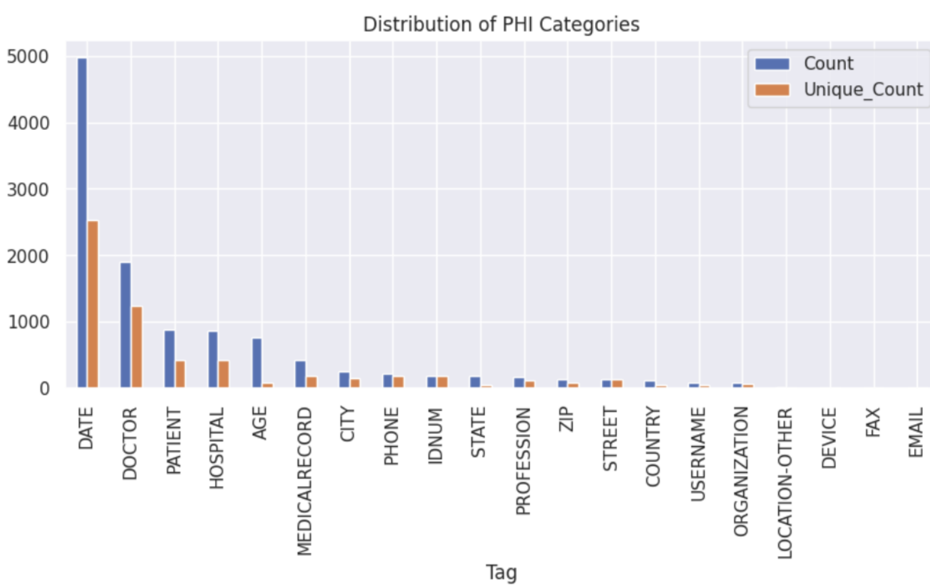


Figure 4: Number of Anonymized Information with and without Duplicates in Each Information Category

There are 20 categories of the PHI in the dataset, including date, doctor name, hospital name, age, medical record number, etc. “Date”, “Doctor Name”, “Patient Name”, “Hospital Name” and “Age” are the frequent categories of private information shown in the medical records, which makes sense as these are the basic required information written in the record. There are also information duplications among different records, such as date, name, and age, where the number of unique values in these categories drops significantly compared to that of all individual values, as shown in Figure 4. The duplicated information can help evaluate if LLM models can recognize similar name entities in different contexts.

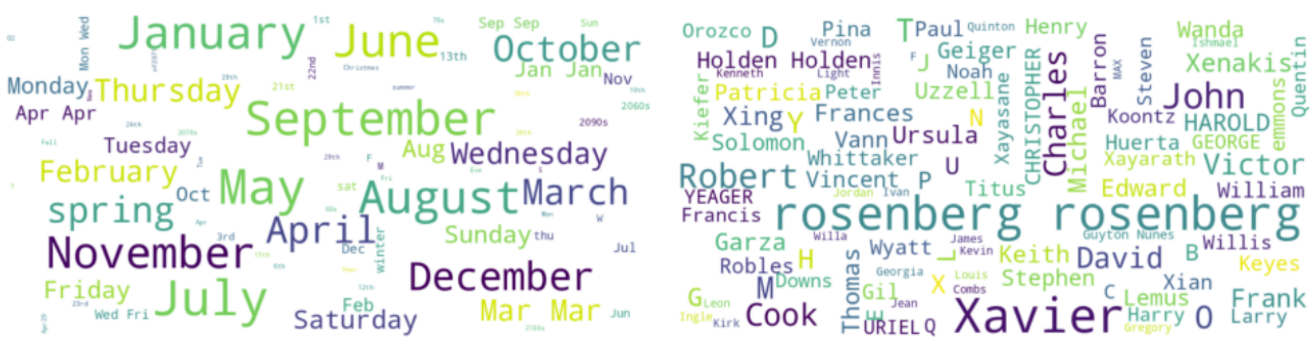


Figure 5: Word Clouds of Private Information in Categories of Date and Doctor

In the categories with the most counts among medical records, which are date, doctor name, patient name, and hospital name, exact texts with high frequency are selected and shown in the word clouds. Several months and names are frequently mentioned, and general terms describing hospitals, like “hospital”, “medical center”, and “clinic” are popularly used in the record as well.

2.2.3 Data Preprocessing

In the n2c2 NLP Research Datasets, each XML file contains <TEXT> and <TAGS> nodes, which save medical annotation texts and PHI tags, respectively. Hence, we extracted the medical annotation text from the <TEXT> node so as to add the original text to the prompts of GPT and Llama2 to perform de-identification later. We also extracted the PHI tags from the <TAGS> node, which represents the true private information to be masked, and will compare them with the results generated by ChatGPT and Llama2 in the further model evaluation part.

We selected the repository of “testing-PHI-Gold-fixed” as each XML file contains the true PHI tags. Due to the computational limit of running billion-parameter LLMs, not all files were selected in testing model performances. We chose combinations of short and long texts: the shortest 100 files, with fewer than 2278 character tokens, and the longest 100 files in different GPT and Llama2 experiments.

The narrative texts in the original ASRS dataset are completely sanitized with annotations, so we filled the masked information back first before performing the de-identification. Specifically, we replaced “Aircraft X” and “Aircraft Y” by randomly sampling from a list of 50 pre-generated aircraft registration numbers that consist of alphabets, numbers, and a slash such as “LC-61A”. We replaced the code “XXX” by sampling one from “C001” to “C030” and replaced “ZZZ” which stands for arriving or departing airports by sampling one from a pre-selected list of 50 IATA airport codes such as “JFK”. In the transfer learning, 50 restored ASRS files were then randomly chosen to evaluate the Llama2 performance on the de-identification task of aviation data.

2.3 Modeling

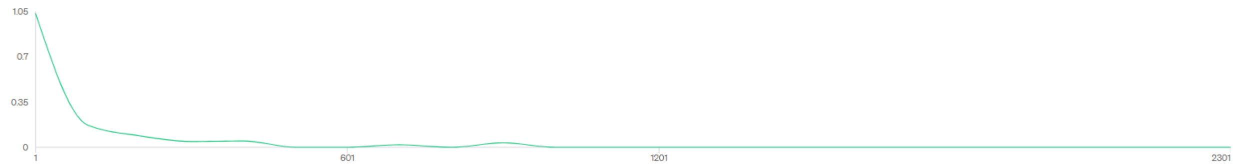
2.3.1 GPT-4

ChatGPT is an advanced large language model (LLM) developed by OpenAI in November 2022. It was designed to generate human-like text and engage in natural language conversations. This model has a wide range of applications, from answering questions to generating content as required, making it a versatile tool for various language-related tasks. In our project, we attempted to utilize both GPT-3.5-turbo and GPT-4 models for the de-identification task as a reference model and applied both models by using the Open AI API in Python, accessing it through the API key. Although GPT-3.5-turbo can perform excellently in anonymizing EHRs by removing PHIs, one of the limitations of ChatGPT-3.5 is that it can only generate at most 4096 tokens in the response, which means it cannot produce the whole anonymized content for some long EHR files in one prompt. Splitting a long clinical record into multiple smaller parts, feeding them into ChatGPT separately through multiple prompts, and combining the responses might be a feasible solution. However, this also causes extra data processing steps and adds more complexity. We also tried GPT-3.5-turbo-16k, which can produce a maximum of 16384 tokens in the response. Nevertheless, it has unexpectedly poor performance and only returns the original text without removing any private information in the response at most times. Compared with GPT-3.5-turbo and GPT-3.5-turbo-16k, GPT-4 outperforms by better identifying PHIs and generating at most 8192 tokens, which can handle most EHR files in only one prompt.

2.3.2 GPT-3.5 Finetune

Although GPT-4 is smarter and has a great improvement in accuracy, it is still insufficient for de-identification tasks. Therefore, we tried to train the GPT model through finetune method to better identify private information in the text. Recently, OpenAI further opened up the usage restrictions of GPT-3.5, allowing users to provide data sets and set parameters, and finetune general GPT-3.5 online.

We compiled and used the training part of n2c2 as a finetune dataset. The training set contains a total of about 300,000 tokens, and the average number of tags required for each sample is 91. In order to eliminate the influence of other aspects, we use exactly the same prompt for training and testing. We uploaded the finetune job for training 3 epochs. The training results are very surprising as Figure 6 shows:



Messages Metrics

Time	Step	Training loss	Validation loss
16:22:59	2301	0.0000	-
16:20:29	2201	0.0000	-
16:17:52	2101	0.0000	-
16:15:13	2001	0.0000	-
16:12:40	1901	0.0000	-
16:10:00	1801	0.0000	-
16:07:27	1701	0.0000	-
16:04:44	1601	0.0000	-
16:02:06	1501	0.0000	-
15:59:25	1401	0.0000	-
15:56:44	1301	0.0000	-
15:54:11	1201	0.0000	-
15:51:31	1101	0.0001	-
15:48:50	1001	0.0000	-
15:46:27	901	0.0346	-
15:43:55	801	0.0000	-
15:41:22	701	0.0192	-
15:38:29	601	0.0003	-
15:35:59	501	0.0002	-
15:33:36	401	0.0491	-
15:31:03	301	0.0442	-
15:28:27	201	0.0889	-
15:25:46	101	0.1717	-
15:23:19	1	1.0352	-

Figure 6: GPT-3.5 Finetuning Result on n2c2 Dataset

We can see from the figure that although we provide enough training data, GPT is fully trained in about 1000 steps. After testing, we found that the model can achieve a completely correct level on shorter texts; even if the text is longer, the model performance only drops slightly, far exceeding the accuracy of human manual classification.

However, Finetuned GPT-3.5 also has some problems. First of all, finetuning does not enable the model to exceed the original input limit. Therefore, for text that is too long (more than 4096 tokens), we still need to split the text, which will affect the final result to a certain extent. As a result, however, we found that OpenAI launched a version with a longer input upper limit (16384 tokens) in November. Using the new model for finetuning will completely solve this problem; secondly, according to OpenAI's policy, Finetuned GPT still needs to pass OpenAI's API to connect and use, and continue to charge. In response to this problem, we hope to solve it in the future by building a local model through distillation learning or finetune llama 2.

2.3.3 Llama2

Llama2 is a large language model released by Meta in July 2023, including Llama Chat and Code Llama. The Code Llama model is fine-tuned to be more suitable for automated code generation to improve programming efficiency. In our project, we use Llama Chat to accomplish private data erasure. To compare the performance of the two models, the prompts we use in each platform for Llama2 and the GPT model used remain identical. As for model deployment, we experimented locally and on the cloud across different platforms to set up the environment. We listed the deploying details and conclusions regarding the model performance and resource use below.

- Google Colab: We tried three approaches: 1) connect directly to Llama2's model using the Hugging face API, 2) download the corresponding transformer for Llama, and 3) a 13B-Chat quantized model.

  Among a variety of API interfaces the Hugging Face provides, we tried the 13B-Chat and 70B models. The 13B-Chat model could get replies but could not mask private information based on our prompts, and the 70B model got a server error response because of too many requests. Moreover, we could not finetune the model connected by API effectively by modifying the prompts. As for the transformer, there was a problem with limited Disk resources when downloading the 70B model, as the Disk limit for Pro members is 150 GB, but the 70B transformer size is over 160 GB.

  We also tried a 13B-Chat quantized model, which can efficiently execute on the T4 GPU and successfully generate anonymized texts based on the prompt. One challenge of using the quantized model is that it limits the total number of input and output tokens to 512, which requires splitting the text into sets of shorter sentences and inputting into the model with multiple runs. However, the major obstacle to adopting the quantized model is the unexpected changes in original text and incomplete generations of results, which may be caused by the limited resources available for this project, resulting in an inaccurate evaluation of the Llama2 model performance.

- HuggingChat: We eventually chose HuggingChat as the platform to use the Llama2 70B-Chat model directly, as there is no limit to the input length, and the text generation is complete and follows the prompt to erase privacy information without the need for finetuning.

Replace private information such as a person's name, date, birthday, gender, location, etc. with #####. Print out the revised text. Text: <![CDATA[Record date: 2069-04-07 Mr. Villegas is seen today. I have not seen him since November. About three weeks ago he stopped his Prednisone on his own because he was gaining weight. He does feel that his shoulders are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his Atenolol for hypertension and 1 Hydroxychloroquine tablet.

Sure, here's the revised text with private information replaced:

Record date: ###### Mr. ###### is seen today. I have not seen him since ##########. About three weeks ago he stopped his ############## on his own because he was gaining weight. He does feel that his ##### are definitely improved. It is unclear what he is actually taking, but I think based on the color of his pills and the timing of the medication that he continues taking his ###################### for hypertension and 1 ###################### tablet.

Figure 7: Llama2 70B-Chat sample result on HuggingChat

For example, from Figure 7, the Llama2 70B-Chat model recognized exactly what the private data is and accurately identified names, symptoms, dates, and other private information in the medical record. Therefore, we chose HuggingChat to implement the Llama2 70B-Chat model. We initially input an implicit prompt together with the original text to be anonymized via chat as user input to get the result. After prompt engineering, we set the explicit prompt as a system prompt and input each text separately via chat to derive de-identified results.

2.3.4 LSTM

In addition to LLM such as GPT and Llama2, we also applied a local deep learning model, Long Short-Term Memory (LSTM), to perform the de-identification task on clinical records, regarding it as a Named Entity Recognition (NER) task. LSTM represents a significant advancement in the field of artificial intelligence, offering a sophisticated approach to address the challenge of long-term dependencies in sequential data processing. We utilized all the training and test files in n2c2 in this part. After splitting each file into a sentence level, we tokenized each word in each sentence and padded shorter sentences to ensure they were equal in length to the longest sentence in our dataset. Then, we used word embeddings to obtain corresponding vector representations to form a complete embedding matrix for further training. According to Richter-Pechanski's research[8], we applied a bi-directional LSTM layer to capture bi-directional long-term dependencies among words to fit the training data. The target of the LSTM is to binarily classify whether each word belongs to PHI or not. We utilized this model to predict the PHI attribute for each word on the whole test dataset and evaluated the results by eliminating additional paddings (details are discussed in Sec 2.6).

2.4 Prompt Engineering and Transfer Learning

Prompt engineering is a technique that intends to figure out a way to effectively communicate with and guide the LLM to produce desired outputs, which involves carefully crafting prompts to elicit specific types of responses or behaviors from the LLM. In order to better guide GPT-4 and Llama2 to anonymize PHI in the clinical records, we formulated both **implicit** and **explicit** prompts, as Figure 8 shows, followed by the original non-anonymized clinical record text. The implicit prompt provides basic instructions to remove private information. In contrast, the explicit prompt offers detailed examples of different kinds of information so that GPT and Llama2 can better realize the de-identification task.

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.

Figure 8. Implicit Prompt v.s. Explicit Prompt for De-identifying EHRs

[8] Richter-Pechanski P, Amr A, Katus HA, Dieterich C, Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports., in Stud Health Technol Inform. 2019 Sep 3;267:101-109, doi:10.3233/SHTI190813.

We also conducted transfer learning by leveraging Llama2 70B Chat to the ASRS dataset to test its compatibility and generality for the de-identification task in a different situation. Given that ASRS has anonymized identifiable information such as aircraft registration number, location, etc., we first de-anonymized by randomly replacing the masked representation with one of the corresponding sampled "real" data as demonstrated in the data preprocessing part. Then, we utilized Llama2 70B Chat to the randomly selected 50 ASRS files to anonymize the three types of information by following the explicit prompt:

> Replace identifiable information in the text with generic placeholders. Specific aircraft registration codes like 'LC-61A', 'AL-Z0X','CO-N4H' should be substituted with 'Aircraft X' and 'Aircraft Y'. Unique airport code, such as 'ATL', 'PEK', 'DXB', etc., must be replaced with 'ZZZ'. Specific codes like 'C001' to 'C030' should be changed to 'XXX'. This approach aims to maintain the original context and meaning of the text while ensuring confidentiality and privacy. The output contains only the changed text.

2.5 Model Evaluation

In order to evaluate to what extent GPT, Llama2, and LSTM models can accurately mask the privacy information in n2c2 and ASRS data, we decided to use typical classification metrics, including accuracy, recall, precision, and F-1 score after defining True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). To compare differences between original texts and anonymized texts rewritten by models, we tokenized both of them by removing extra white spaces and punctuations. In the original text, we divided tokens into two groups: tokens containing privacy information and tokens with general information, which should not be masked. In the n2c2 dataset, tokens with privacy information are already extracted as PHI tags. In the ASRS dataset, tokens to be masked are annotations like "XXX" and "Aircraft Y". Then we defined TP, TN, FP, and FN as follows:

- **TP:** the number of tokens that should be masked and are successfully masked. We computed TP by checking the number of tokens with privacy information that do not appear in the anonymized text.
- **TN:** the number of tokens that should not be masked and are kept as original. We computed TN by checking the number of tokens with general information that appear in the anonymized text.

- **FP:** the number of tokens that should be kept as original but are masked. We computed FP by checking the number of tokens with general information that is not kept in the anonymized text.
- **FN:** the number of tokens that should be masked but are not successfully masked. We computed FN by checking the number of tokens with privacy information that are kept in the anonymized text.

With the above terms, we computed metrics to evaluate model performance on private information de-identification by following formula and calculated the average scores of these metrics over all samples:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

$$F1\ score = \frac{2 * precision * recall}{precision+recall}$$

2.6 Results

The performance of LLM models and LSTM with different prompts is evaluated on the n2c2 medical record testing sets, and transfer learning of Llama2 is applied using ASRS aviation data, as shown in Table 1. LSTM was implemented as a baseline to compare LLMs with traditional deep learning methods. As for testing files for each method, the n2c2 50 short texts and 100 short texts refer to the first 50 and first 100 medical records with the smallest number of tokens, and 100 long texts refer to the first 100 medical records with the largest number of tokens.

The evaluation result revealed that GPT-4 with explicit prompts outperformed the one with implicit prompts across all metrics, notably in the recall, which is essential for complete data anonymization. Fine-tuning GPT-3.5 further enhanced its performance, achieving near-perfect scores in all metrics. Also, GPT-3.5 with finetuning performed excellently in terms of both short and long texts, therefore indicating its superior effectiveness in completely identifying and correctly anonymizing private information.

Llama2 70B-Chat demonstrated better performance in de-identifying private information with implicit prompts compared to explicit prompts with significantly higher precision. 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, proving the successful application of transfer learning by tuning the prompts based on the dataset.

Method	Prompt	Accuracy	Recall	Precision	F-1 Score	Testing Files
LSTM	—	91.01%	77.73%	63.99%	70.19%	n2c2 all texts
GPT-4	Implicit	96.44%	75.94%	73.63%	74.77%	n2c2 100 short texts
GPT-4	Explicit	97.44%	95.03%	74.81%	83.72%	n2c2 100 short texts
GPT-3.5 fine-tuned (Short Text)	Explicit	99.97%	99.27%	100.00%	99.61%	n2c2 50 short texts
GPT-3.5 fine-tuned (Long Text)	Explicit	99.88%	98.57%	98.99%	98.74%	n2c2 100 long texts
Llama2 70B-Chat	Implicit	89.48%	59.20%	46.00%	45.80%	n2c2 50 short texts
Llama2 70B-Chat	Explicit	85.93%	58.01%	27.08%	36.92%	n2c2 100 short texts
Llama2 70B-Chat (Transfer Learning)	Explicit	96.47%	72.53%	12.38%	21.15%	ASRS random 50 files

Table 1. Model Performance on Medical Record (n2c2) De-identification and Transfer Learning on Aviation Data (ASRS)

In terms of model deployment, there are both advantages and disadvantages for GPT and Llama2 models in the context of our task and experiments so far. GPT is easy to deploy with fewer computation resources, compared to all Llama2 models, and has a larger token input. However, it may take a longer time to produce results compared to Llama2 70B-chat. Llama2 can more efficiently use resources with a smaller parameter model. However, there is a large resource needed for deployment, especially with a bigger parameter model. Also, there is a limit on the length of input tokens, which may require additional splits of text for the model input.

In 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. Due to limited computing resources available for this project, Llama2's performance falls short in de-identification tasks compared to GPT-4 and fine-tuned GPT-3.5 models.

3. Discussion

This project has achieved notable success in developing and refining an Information Privacy Control mechanism by utilizing Large Language Models, particularly Llama2 and GPT, to bolster data privacy within corporate frameworks. Through extensive comparative analyses and meticulous fine-tuning, these models have demonstrated their proficiency in de-identifying sensitive data, thereby ensuring compliance with privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA). The application of these models across a spectrum of datasets, including those from the healthcare and aviation sectors, has further highlighted their adaptability and the potential for expansive application.

Take Home Messages:
1. **Enhanced Data Privacy through Large Language Models**: The research underscores that Large Language Models, when optimally fine-tuned, can substantially enhance data privacy and regulatory compliance across a diverse range of industries.
2. **Importance of Continuous Refinement and Ethical Oversight**: The findings emphasize the necessity of ongoing refinement and the incorporation of ethical considerations for the responsible implementation of these models in real-world scenarios.

3. **Foundation for Future Innovations**: This research lays a significant groundwork for future advancements, signaling a promising direction in the integration of sophisticated AI technologies to fortify data security and privacy in corporate environments.

For future work, our primary goal is to allocate more resources towards deploying the models, especially Llama2, delving deeper into its algorithm to identify and implement more effective fine-tuning strategies. Also, prompt engineering for LLMs often involves a trial-and-error process, requiring precise composition for specific tasks, so we would also focus on enhancing system prompts. Additionally, we plan to explore other models that offer unique features or improved performance for specific tasks. A key focus will be on enhancing dataset diversity. Incorporating a wider variety of datasets, particularly those from different domains or with varied data structures, is crucial for ensuring the robustness and generalizability of our models. This diversity is essential for preparing the models to handle a range of real-world scenarios effectively.

Addressing ethical considerations and mitigating potential biases in LLMs is also a priority as these models become more integrated into data privacy tasks. The potential ethical concerns of utilizing LLMs include the risk of unintended misuse of the technology, privacy violations, and the amplification of biases present in the training data. Ensuring that the data used for training and fine-tuning is representative and unbiased is crucial for the ethical application of these technologies. Collaboration and cross-disciplinary research will also play a significant role in our future work. Engaging with experts in fields like data security, legal studies, and ethics will provide holistic insights and innovative solutions.

Additionally, piloting the models in real-world settings and establishing feedback loops will be essential for continuous improvement based on practical insights and challenges. Customizing these models for specific industries, such as healthcare, finance, or aviation, is another critical area of focus, considering the varying data privacy needs across different sectors. Furthermore, as datasets grow and the complexity of tasks increases, enhancing the scalability and computational efficiency of these models will be imperative.

By pursuing these avenues, we aim to not only build upon the existing foundations but also push the boundaries in the application of LLMs for data privacy and protection, creating more sophisticated, ethical, and efficient solutions for the challenges ahead.

4. Contribution

- **Jiawen Song:** Responsible for communicating with TAs. Main contributor to research and literature review of Llama2 models; tested and compared said models
- **Kaiyuan Liu:** Main contributor on applying GPT-4 and LSTM to de-identify n2c2 dataset and formulating the free-text result evaluation function.
- **Longcong Xu:** Deployed the LLama2 model locally, on GCP, on Google Colab, and prompt-engineered the LLama2 model on each platform.
- **Shaonan Wang:** Set up meetings with mentor. Main contributor to exploratory data analysis, Llama2 deployments, and organizing Llama2 results on n2c2 and ASRS data.
- **Shukai Wang:** Main contributor on applying GPT-4 and fine-tuning GPT-3.5 to de-identify n2c2 dataset and responsible for organizing and processing dataset.