

Information Privacy Controls for Corporate Use of Public LLMs

by

Longcong Xu (lx2305)

Shukai Wang (sw3715)

Shaonan Wang (sw3720)

Jiawen Song (js6123)

Kaiyuan Liu (kl3447)

First Progress Report

Working with GE Aviation and
Data Science Institute at Columbia University

October 28, 2023

Table of Contents

Contents	Page #
1. Introduction.....	3
2. Background and Data.....	4
2.1 Background.....	4
2.2 Data.....	5
3. Exploratory Data Analysis.....	7
4. Modeling.....	10
4.1 Data Preprocessing.....	10
4.2 ChatGPT.....	10
4.3 Llama2.....	11
4.4 Model Evaluation.....	14
5. Conclusion.....	17
6. Next Steps.....	19

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 and advance algorithms tailored to facilitate the corporate integration of publicly available Large Language Models (LLMs). Given the existing constraints of conventional privacy techniques and the distinctive LLM requisites within corporate settings, our project primarily aims to use and finetune the Llama2 model to aid corporations in applying LLMs for the protection and assurance of data privacy.

Leading up to the initial checkpoint in this project, our team's primary focus lies in the utilization of the Llama2 and ChatGPT models in the context of the 2014 i2b2/UTHealth de-identification challenge dataset. Our objective is to formulate appropriate prompts for removing protected health information (PHI) in compliance with the Health Insurance Portability and Accountability Act (HIPAA). Following this phase, we will conduct a comprehensive comparative assessment of the performance of Llama2 and ChatGPT. Subsequent steps will be centered around refining and fine-tuning Llama2 to enhance its effectiveness. We also plan to perform transfer learning by applying ChatGPT or fine-tuned Llama2 models to the ASRS dataset, which reports the aviation safety narratives, after manually filling in the masked information.

The first stage of this project comprises three key steps. First, we deployed the Llama2 environment locally and on the Google Colab and Google Cloud Platform (GCP) to select the most suitable modeling environment, considering the resource limits. The second step involved the exploration of suitable "deidentification" prompts for application to the Llama2 and ChatGPT models. We read the XML files from the 2014 i2b2 dataset, integrated prompts into both models and generated masked text for each file. The third step entailed further processing of the generated text and a comparative analysis with the pre-existing masked files to evaluate the performance of both models.

2. Background and Data

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 the US Health Insurance Portability and Accountability Act (HIPAA), all the identifying information in patient 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).

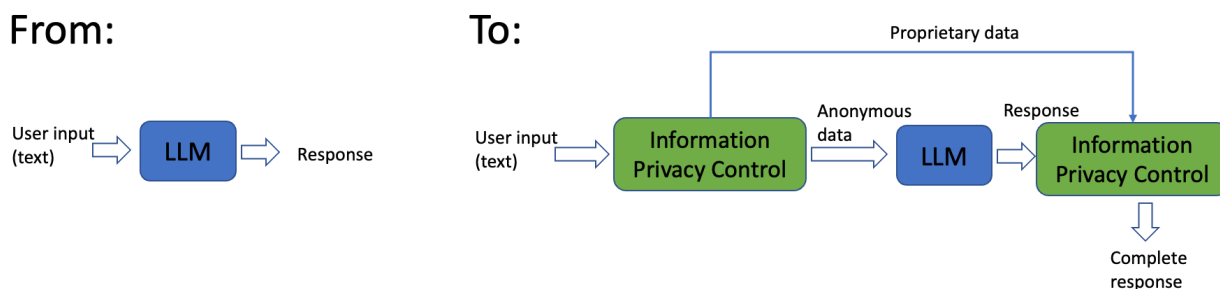


Figure 1: Definition of Information De-identification

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 data¹. With the advent of Large Language Models (LLMs) in 2023, there is a growing interest in exploring LLM-based de-identification methods and utilizing these models for anonymizing clinical records². Among all the popular LLMs, ChatGPT and Llamas 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 ChatGPT and Llama2, which could pave the way for subsequent fine-tuning.

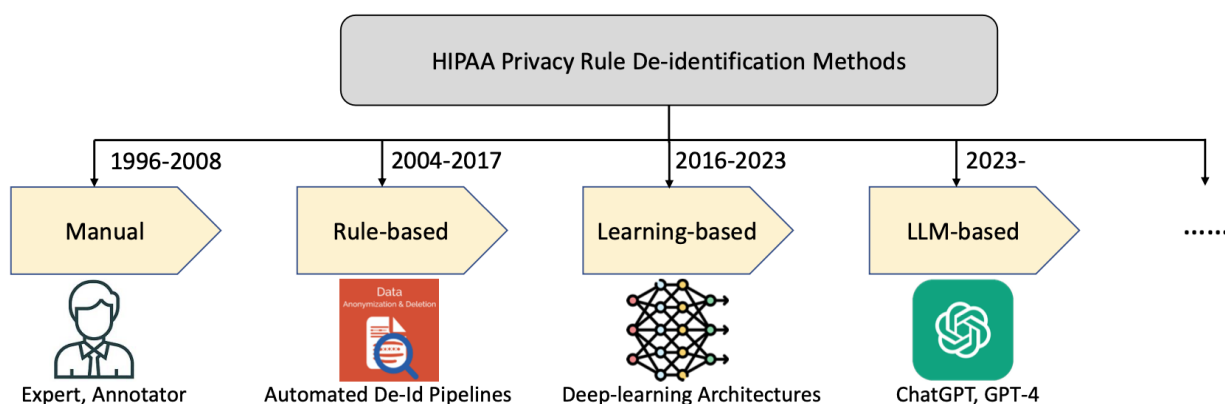


Figure 2: History of De-identification Methods in Accordance with HIPAA

2.2 Data

We will test the performance of ChatGPT and LLama2 models on the testing set of n2c2 NLP Research Data Sets. After we fine-tuned the LLama2 model, we will apply transfer learning and test the performance of our fine-tuned model on the ASRS database.

The data used for this first progress project is 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

¹ 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.

² 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.

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 2014 i2b2 de-identification track, aiming to remove protected health information (PHI) in the medical record. It was later used by 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](https://n2c2.dbmi.hms.harvard.edu/data-sets)³.

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 sanitized by notations like XXX and ZZZ. 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.⁴

³ “Data Sets.” *National NLP Clinical Challenges (n2c2)*, <https://n2c2.dbmi.hms.harvard.edu/data-sets>. Accessed 27 October 2023.

⁴ ASRS Database Online - Aviation Safety Reporting System.” *NASA ASRS*, <https://asrs.arc.nasa.gov/search/database.html>. Accessed 27 October 2023.

3. Exploratory Data Analysis

In the first stage, the testing set in the n2c2 dataset is 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.

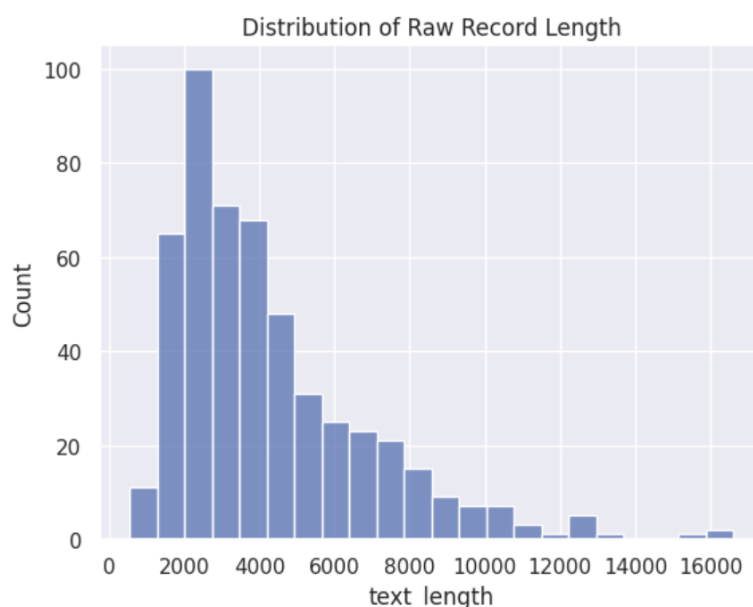
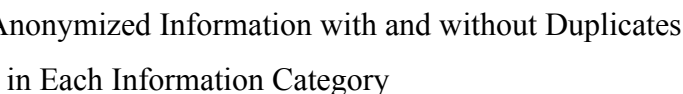


Figure 3: Histogram of Medical Record Lengths

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.



A word cloud featuring names and dates. The names are arranged in a circular pattern, with 'rosenberg' being the most prominent. The dates are arranged in a rectangular pattern, with 'January' and 'June' being the most prominent.

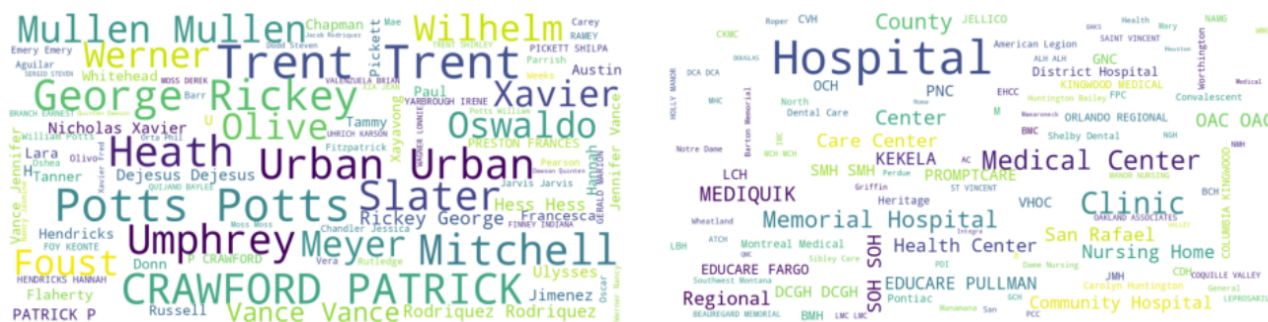


Figure 5: Word Clouds of Private Information in Categories of Date, Doctor, Patient, and Hospital separately (from left to right, top to bottom)

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.

4. Modeling

This section summarizes the procedure of text data processing, environment setting-up, anonymous data generation using ChatGPT and Llama, and model performance evaluation.

4.1 Data Preprocessing

As mentioned in Section 2, the first progress report focuses on the n2c2 NLP Research Datasets, a series of XML files storing electronic health records (EHRs), clinical notes, radiology reports, pathology reports, etc. Each XML file contains <TEXT> and <TAGS> nodes, which saves 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 ChatGPT 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.

For the first progress report, 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, we selected the 50 XML files whose length of medical annotation text is smaller than 1780 words for the initial results, including “311-02.xml” and “165-02.xml”, etc.

4.2 ChatGPT

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, 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 back 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. Therefore, we decided to focus on **GPT-4** for the de-identification task of 50 files in the first stage.

After initializing GPT-4 through the Open AI API, we tested it by using different prompts to trigger the responses. We finally specified the prompt for the de-identification task using GPT-4 as follows, followed by the original non-anonymized clinical record text:

- “Replace private information such as a person’s name, date, birthday, gender, location, etc. with #####. Print out the revised text.” + [ORIGINAL_TEXT]

4.3 Llama2

Llama2 is a large language model released by Meta in July 2023, including Llama Chat and Code Llama. 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.

- Local: We downloaded the Llama2 7B, 7B-Chat, 13B, and 13B-Chat models locally. As the original models need to consume too much memory, the local memory is difficult to support. We first quantized the models by reducing the size of each model to under 10GB. After activating Chatbot locally, we tested each of the four lightweight models. However, after the models are quantized, the restrictions on the input and output tokens become stricter. In the original model, the input limit of the model token should be 4000, but the local model only accepts 508 token input. Therefore, we made a small number of attempts to enter statements.

```
User:Replace private information such as a person's name, date, birthday, gender, location, etc. with ####. Print out the revised text: 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.
Bob: Sure. The largest city in Europe is ####, the capital of Russia.

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.
User:Replace private information such as a person's name, date, birthday, gender, location, etc. with ####. Print out the revised text: 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.

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.
```

Figure 6: Local Llama2 13B-Chat Sample Results

Even though we use the largest 13b-chat model locally, the output is still unsatisfactory. So we wanted to move to the cloud platform for more experimentation.

- Google Colab: On Google Colab, we tried two approaches. The first was to connect directly to Llama2's model using the Hugging face API. Second, download the corresponding transformer for Llama. Hugging Face provides a variety of API interfaces. We tried the 13B-Chat and 70B models, and the 13B-Chat model could get replies but could not mask private information. Moreover, through the API connection model, we could not finetune it effectively by modifying the prompts. The 70B model got a server error response because of too many requests.

```
[4] import requests

API_URL = "https://api-inference.huggingface.co/models/meta-llama/Llama-2-13b-chat-hf"
headers = {"Authorization": "Bearer "

def query(payload):
    response = requests.post(API_URL, headers=headers, json=payload)
    return response.json()

output = query({
    "inputs": "Replace private information such as a person's name, date, birthday, gender, location, etc. with #####. 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. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the car..."

[7] output[0]['generated_text']
```

'Replace private information such as a person's name, date, birthday, gender, location, etc. with #####. 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. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the car...'

Figure 7: API Calls Llama2 13B Chat on Google Colab

To meet the computational resources of large models, we upgraded Colab to pro membership. With the support of the NVIDIA T4 GPU, we tried to download the transformer for better testing.

```
[13] sequences = pipeline(
    "Remove private information such as a person's name, date, birthday, gender, location, etc.. 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. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the car..."
    do_sample=True,
    top_k=10,
    num_return_sequences=1,
    eos_token_id=tokenizer.eos_token_id,
    max_length=200,
)

[15] for seq in sequences:
    print(f"Result: {seq['generated_text']}")
```

[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. He is concerned because of the relatively recent onset of difficulties turning his head to the right. When he does this, he will note that he feels as though he is going to pass out although this has not actually happened. This only occurs when he turns to the right and not to the left. He has no visual changes otherwise and denies any headache or other cranial complaints. On examination today, BP 120/80. He has no bruits over the car...'

Figure 8: Llama2 13B Chat transformer on Google Colab

We can see that before finetune, Llama2 13B-Chat will not erase private data according to our instructions. When we downloaded the 70B model, we ran into the problem of not having enough Disk resources; the Disk limit for Pro members is 150 GB, but the 70B transformer size is over 160 GB.

- GCP: Vertex AI provides a platform for deploying model pipelines on GCP. We first created an instance for storing the model, and for GPU, we stayed consistent with Google Colab usage and chose NVIDIA T4. Due to insufficient credits (we redeemed \$200 in our

billing account), the credits were used up after the model was downloaded, and after a few prompts were attempted, there was an insufficient quotas issue.

- HuggingChat: HuggingChat is the only place we've come across where we can use the Llama 2 70B-Chat model directly. the 70B-Chat model has done a good job of following our instructions for privacy erasure without the need for finetune.

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 9: LLama2 70B Chat sample result on HuggingChat

As we can see, the Llama2 70B-chat model recognizes exactly what the private data is and can accurately look for people's names, symptoms, dates, and more in the text. The only downside is that it doesn't mask out Mr. and retains the gender information. **Thus we use HuggingChat to generate the results and compare them with the GPT results.**

4.4 Model Evaluation

In order to evaluate to what extent GPT-4 and Llama2 can accurately mask the PHIs, we decided to use typical classification metrics, including accuracy, recall, precision, and F-1 score. Since

we are processing text data, we need to define True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) in the context of de-identifying clinical records. Before that, we extracted original medical annotation data and true tags and tokenized both of them with removing extra white spaces. By feeding the non-anonymized clinical record, we obtained the rewritten anonymous text data and tokenized it with removing extra white spaces. We also identified non-PHIs that should not be masked by separating non-tag tokens from the original text. Then we defined TP, TN, FP, FN as follows:

- **TP:** the number of tokens that should be masked and are successfully masked. We computed TP by checking the number of tags that do not appear in the returned anonymous data.
- **TN:** the number of tokens that should not be masked and are not masked. We computed TN by checking the number of non-PHIs that appear in the returned anonymous data.
- **FP:** the number of tokens that should not be masked but are successfully masked. We computed FP by checking the number of non-PHIs that do not appear in the returned anonymous data.
- **FN:** the number of tokens that should be masked but are not successfully masked. We computed FN by checking the number of tags that appear in the returned anonymous data.

With the above four items, we compute the metrics for each XML file by following formulas:

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

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

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

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

We further calculate the average values of the above four metrics over the 50 samples, and the results will be discussed in Section 5.

Considering that Llama2 70B-chat is unable to produce the whole anonymous data for long EHRs due to the unsatisfying memory mechanism, we also set up a rule where we will not consider the files whose length of original clinical record differs the length of anonymized returned response by over 10% for the final average metrics. No file that meets the anomaly condition exists for the 50 samples. We also provided a split function that splits a long text into multiple smaller parts given the customized number of tokens in a slice to handle the situation where Llama2 70B-chat cannot generate the whole response once.

5. Conclusion

In the first stage, we clearly defined the scope of the project to develop an Information Privacy Control mechanism using LLM models, specifically ChatGPT and Llama2, in the area of healthcare and aviation. The task is to identify and sanitize the privacy information, such as names and dates, in the medical records and aviation safety narratives. We first use the pre-trained ChatGPT and Llama2 models to anonymize the testing set in the n2c2 clinical dataset. As shown in Table 1, based on the metrics defined in Section 4.4, we compare the model performance between these two models. Both models show a high accuracy on average. ChatGPT shows better results in accuracy and F1 than Llama2, with a difference of 5% and 14% respectively, showing that ChatGPT performs better to correctly identify the privacy information from a general perspective. ChatGPT displays a huge advantage on the Llama2 in terms of precision, meaning among all information identified, it has a higher correctness. Llama2 outperforms ChatGPT in the recall value, meaning it can better identify the correct privacy information in the entire text.

	ChatGPT	Llama2
Average Accuracy	0.9461	0.8948
Average Recall	0.5504	0.5920
Average Precision	0.7094	0.4600
Average F-1	0.5962	0.4580
Pros	<ul style="list-style-type: none"> - Easy to deploy with less computation resources, compared to all Llama2 models - Larger token input 	<ul style="list-style-type: none"> - More efficient use of resources with smaller parameter model
Cons	<ul style="list-style-type: none"> - Longer time to produce results compared to Llama2 70B-chat 	<ul style="list-style-type: none"> - Large resource needed for deployment, especially with bigger parameter model

		<ul style="list-style-type: none"> - Limited length of input tokens - Less satisfying model result
--	--	--

Table 1. Performance of ChatGPT and Llama2 on n2c2 Testing Set and Model Comparison

In the second stage, we will set these results as the baseline to compare with the model performance on the same dataset after fine-tuning the model. We also compare the advantages and disadvantages of the two models in the context of our task and experiments so far.

6. Next Steps

For future work and next steps, we have several plans:

1. Grasping the GPT-ner Model:

We aim to delve into research articles detailing the GPT-ner model's structure and abilities.

Based on preliminary tests, we will determine our reliance on this model.

2. Preparing the Dataset:

For Fine-tuning: Our objective is to collect and label additional data similar to the 2014 i2b2 dataset.

For Transfer Learning: Our goal is to obtain the ASRS dataset, pinpointing and marking sensitive data points that require anonymization.

3. Fine-tuning the Model:

We intend to refine the Llama2 13b model, optimizing it for superior de-identification task performance.

4. Learning through Data Transfer:

Our strategy involves initiating with the weights of the refined model, and then tailoring the model to suit the ASRS dataset, with a keen focus on detecting and masking sensitive information.