De-identification of PHI in Electronics Health Records using LM's

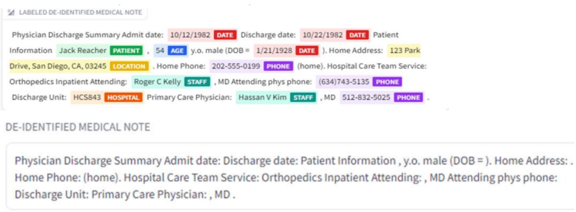
Anveshika Kamble, Akshat Dhamale, Dilip Teja

1 Introduction

The most important data stored in hospitals and healthcare systems daily are the electronic health records (EHRs), which include data used to identify health concerns, enhance patient treatment, and support public policy efforts (Kruse et al., 2016). Despite its significance, the privacy issues around the disclosure of Protected Health Information (PHI) are a significant barrier. Under the Health Insurance Portability and Accountability Act (HIPAA) in the United States (Annas, 2003), before a record may be made more publicly available, 18 direct identifiers must be eliminated or safeguarded. De-identifying PHI is a laborious process, and the scope of current tools might be restrictive (Hripcsak and Albers, 2013).
Our approach involves leveraging pre-trained Large Language Models (LLMs), such as BERT, BioBERT, RoBERTa, ELECTRA, and LLama 3.2. Incorporating domain-specific pre-training and advanced fine-tuning techniques, their performances were evaluated on the PHI de-identification tasks using the i2b2(Informatics for Integrating Biology and the Bedside) dataset. An extensive comparative analysis is performed considering all different parameters. The analysis results show that Llama is the most suitable model for the PHI de-identification task, achieving the highest F1-scores.

2 Project Motivation

Manual de-identification is quite demanding because it takes a lot of time and labor. For annotations by the initial annotator, the median cost per individual PHI can be $0.71, but for annotations by the fourth annotator (4th person to look for PHI), it can be $377 (Carrell et al., 2016; Kovačević et al., 2024). Consequently, it has been necessary to use natural language processing (NLP) to automate and reduce the cost of this operation. De-identification is usually modeled as a named entity recognition (NER) task. NER is a subtask of Information Extraction, which aims to identify and classify named entities within unstructured text data. It can be carried out in two ways - removal and substitution. Removal includes detecting PHI tags and completely omitting them whereas substitution implies detecting PHI tags and replacing them with synthetic data. Our primary goal is to accurately detect PHI tags and omit them completely from the EHR which ensures patient safety. An example of medical de-identification is shown in Figure 1.



LABELED DE-IDENTIFIED MEDICAL NOTE

Physician Discharge Summary Admit date: 10/12/1982 DATE Discharge date: 10/22/1982 DATE Patient Information Jack Reacher PATIENT , 54 AGE y.o. male (DOB = 1/21/1928 DATE). Home Address: 123 Park Drive, San Diego, CA, 03245 LOCATION . Home Phone: 202-555-0199 PHONE (home). Hospital Care Team Service: Orthopedics Inpatient Attending: Roger C Kelly STAFF , MD Attending phys phone: (634)743-5135 PHONE Discharge Unit: HCS843 HOSPITAL Primary Care Physician: Hassan V Kim STAFF , MD 512-832-5025 PHONE .

DE-IDENTIFIED MEDICAL NOTE

Physician Discharge Summary Admit date: Discharge date: Patient Information , y.o. male (DOB =). Home Address: . Home Phone: (home). Hospital Care Team Service: Orthopedics Inpatient Attending: , MD Attending phys phone: Discharge Unit: Primary Care Physician: , MD .

Figure 1: Medical de-identification example from i2b2 dataset

3 Literature Review

Over the years, several methods have been developed to automate the process of de-identification of PHI tags in EHRs, ranging from rule-based systems to the more recent advances in machine learning, especially large language models (LLMs).

3.1 Earlier methods for NER detection

Early NER methods for PHI identification relied heavily on rule-based approaches or feature-engineered machine learning models. These relied heavily on manual feature extraction, where common features included part-of-speech tags, word boundaries, and gazetteer lists for common PHI terms. These had limitations in handling the complexity and variability inherent in clinical narratives, especially in identifying PHI in unstructured or noisy data.

Deep learning models, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), became widely adopted for NER tasks. These models were capable of capturing sequential dependencies in text, which is essential for understanding context and relationships between entities in clinical notes. (Liu et al., 2017) demonstrated that a combination of LSTMs and CRFs could significantly improve the identification of PHI in clinical records achieving overall precision score of 95% when applied to the i2b2 dataset.

3.2 Transformer based models for PHI De-identification

Transformer-based architectures have significantly advanced PHI de-identification by leveraging contextual embeddings and attention mechanisms. These models have consistently outperformed traditional rule-based or statistical methods in recognizing and anonymizing sensitive data in clinical texts.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin, 2018) and its variants, such as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019), have been foundational in PHI de-identification. BERT achieves F1 scores of over 94% on the i2b2 2014 dataset by leveraging bidirectional context. BioBERT, pre-trained on PubMed and PMC data, enhances this performance, reaching up to 95.6% F1 scores for PHI recognition tasks. RoBERTa (Robustly Optimized BERT Pretraining) (Liu, 2019) outperforms BERT through better training optimization techniques, achieving F1 scores of 96% on PHI deidentification datasets. ELECTRA (Clark, 2020) further improves computational efficiency and maintains comparable performance, achieving an F1 score of approximately 95.4% by focusing on discriminative learning.

Domain-specific models like ClinicalBERT and KeBioLM (Yuan et al., 2021) achieve even higher accuracy for clinical text. ClinicalBERT, fine-tuned on MIMIC-III, achieves F1 scores of up to 96.5% on de-identification tasks . KeBioLM integrates knowledge-based embeddings, achieving F1 scores of 97% for nested and complex entities, showcasing its adaptability to intricate PHI categories. Additionally, BioBART (Yuan et al., 2022), a pre-trained biomedical transformer model based on BART (Bidirectional and Auto-Regressive Transformers), achieves state-of-the-art performance for both entity recognition and text summarization in the biomedical domain, with competitive F1 scores of 96.8% on MIMIC dataset.

3.3 Large Language Models (LLMs) for PHI De-identification and NER tasks

The rise of large language models (LLMs) has further pushed the boundaries of de-identification tasks by leveraging extensive pretraining on diverse datasets and task-specific fine-tuning.

MedPaLM-2 (Singhal et al., 2023) and DeID-GPT (Liu et al., 2023) are among the most prominent models in this domain. MedPaLM-2, fine-tuned on medical texts, achieves F1 scores of 96.2% by utilizing task-specific enhancements such as better pretraining corpora and fine-grained token-level recognition. DeID-GPT, specifically developed for de-identification, surpasses MedPaLM-2 with an F1 score of 97.1%, attributed to its prompt-based approach for handling sensitive entities. GatorTron (Yang et al., 2022), a transformer designed for medical applications, excels in multi-task NLP, including semantic similarity and NER. On the i2b2 dataset, it achieves an F1 score of 96.8%, demonstrating its robustness for clinical text processing. Similarly, BioGPT, a biomedical GPT model, performs well in NER and relation extraction tasks, showcasing versatility in multiple downstream applications with an F1 score of 96.5%.

ClinicalT5 (Lu et al., 2022), an encoder-decoder architecture, has been shown to excel in classification and de-identification tasks, achieving F1 scores of 96% on i2b2, benefiting from its capability to capture both contextual and sequential dependencies effectively. Recent domain-specific adaptations of popular LLM architectures like GPT and LLama have further elevated their performance. For instance, MedGPT(Kraljevic et al., 2021) adapts GPT-3 for biomedical tasks, achieving state-of-the-art results in PHI redaction. Similarly, BioLLaMA, a fine-tuned variant of LLaMA-2, focuses on tasks like PHI de-identification and clinical summarization, with F1 scores of over 96% on the i2b2 dataset.

4 Data

Here, we are using data from Informatics for Integrating Biology and the Bedside (i2b2) that includes unstructured clinical narratives. The dataset comprises 1304 longitudinal medical records from 296 diabetic patients from different U.S. hospitals;

790 of these records are utilized for training, while the remaining 514 are used for testing (Kohane et al., 2012). A portion of the 514 testing files represents our development data collection. The distribution of PHI categories in the whole corpus is visualized in Figure 2 and 3.

The corpus was annotated by licensed professionals following i2b2's regulations. The PHI categories were divided into 24 subcategories and 7 major categories. With tags defining the text's beginning and ending locations, the annotation is captured in XML format (e.g NAME start=12, end=18 Text='Dr. John' TYPE='DOCTOR').

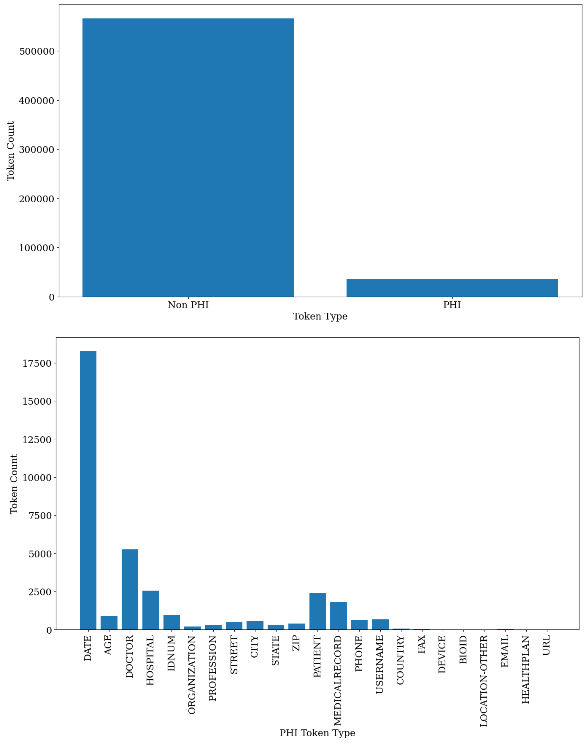


Figure 2: PHI categories occurrence distribution

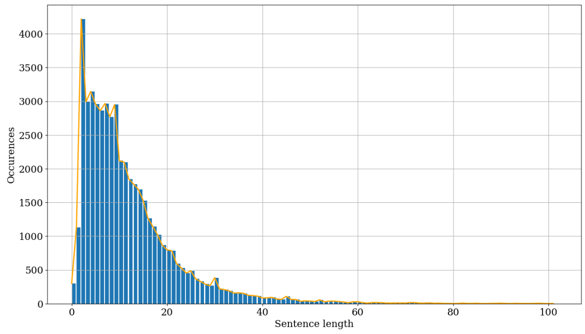


Figure 3: Sentence length occurence distribution

4.1 Data preprocessing

Most data processing focuses on transforming the raw XML file into a format that LLM's can use. The data began as entirely unstructured medical notes, thus we needed to spend a significant amount of time in processing the data.

4.1.1 Extraction

We first extract the tags from "TAGS" attribute and clinical note from the "TEXT" attribute. Clinical note is raw and unstructured text. This text is divided into sentences using REGEX by detecting new lines and spaces.

To address the imbalance in the dataset (PHI and non-PHI tags), we filter sentences which have PHI tags associated with them. The sentences which don't have PHI tags associated are avoided.

4.1.2 Tokenization

The sentences and associated tagged words are tokenized using BERT tokenizer for BERT based models and LLama tokenizer for LLama 3.2 model. Each sentence is converted into list of tokens extracted by the tokenizer.Eg. 'Date:01-25-2012' gets broken down into tokens 'Date',':','01-25-2012'.

4.1.3 BIO encoding

BIO (Begin-Inside-Outside) encoding is a labelling scheme commonly used in NLP tasks, especially for NER tasks like this one. Tokenized PHI entity is labelled in the following way

1. B- (Begin) : Indicates the beginning of an entity

2. I- (Inside): Indicates a token that is part of an entity or chunk but is not the first token.

3. O- (Outside): Indicates a token that is not part of any entity or chunk.

4.1.4 Seperators

We made sure that every sentence have CLS token in the beginning of the sentence and SEP at the end of the sentence (eg. [CLS, token ... , SEP] to help separate the text.

4.1.5 Padding

Each sentence have different length of tokens. We padded every sentence to same length which the BERT model accepts (eg. max-length = 20). Since, it's the multi-class base model, adding special tokens ([CLS] at the start and [SEP] at the end) as 24th and 25th class. It pads the sentences to ensure consistent length across all inputs.

Table 1: Count/percentage of PHI entities in each of the train, validation and test samples of i2b2 corpus

	Training Sample Count (Percent)	**Validation Sample Count (Percent)**	**Test Sample Count (Percent)**
NON-PHI	396674 (94.036)	56118 (94.064)	113965 (94.088)
DATE	12766 (3.0263)	1804 (3.0238)	3687 (3.0439)
DOCTOR	3689 (0.8745)	526 (0.8816)	1050 (0.8668)
HOSPITAL	1794 (0.4252)	253 (0.4240)	510 (0.4210)
PATIENT	1703 (0.4037)	230 (0.3855)	464 (0.3830)
AGE	622 (0.1474)	98 (0.1642)	185 (0.1527)
MEDICALRECORD	1242 (0.2944)	159 (0.2665)	407 (0.3360)
CITY	393 (0.0931)	59 (0.0988)	106 (0.0875)
STATE	182 (0.0431)	25 (0.0419)	63 (0.0520)
PHONE	477 (0.1130)	75 (0.1257)	101 (0.0833)
USERNAME	474 (0.1123)	62 (0.1039)	125 (0.1031)
IDNUM	669 (0.1585)	91 (0.1525)	187 (0.1543)
PROFESSION	225 (0.0533)	30 (0.0502)	57 (0.0470)
STREET	374 (0.0886)	57 (0.0955)	85 (0.0701)
ZIP	279 (0.0661)	37 (0.0620)	68 (0.0561)
ORGANIZATION	152 (0.0360)	20 (0.0335)	34 (0.0280)
COUNTRY	39 (0.0092)	8 (0.0134)	12 (0.0099)
FAX	21 (0.0049)	2 (0.0033)	8 (0.0066)
DEVICE	14 (0.0033)	2 (0.0033)	3 (0.0024)
EMAIL	19 (0.0045)	2 (0.0033)	5 (0.0041)
LOCATION-OTHER	11 (0.0026)	0 (0.0)	1 (0.0008)
URL	4 (0.0009)	1 (0.0016)	1 (0.0008)
HEALTHPLAN	2 (0.0004)	0 (0.0)	0 (0.0)
BIOID	3 (0.0007)	0 (0.0)	1 (0.0008)

5 Methods

The methodological pipeline can be visualized in Figure 4. The approach is organized into three primary phases: data processing, model training, and performance evaluation. Raw text records from i2b2 dataset are first segmented into individual sentences to streamline training and evaluation. Following this, PHI tags such as NAME, DATE, and AGE are mapped to the corresponding sensitive information in the text, marking them for de-identification. Tokenization for the respective model is then applied, where the raw sentences are converted into sequences of tokens compatible with the input requirements of the language models. BIO (Begin-Inside-Outside) encoding is employed to represent PHI tags at the token level, ensuring granularity and precision in tagging. Additionally, end-of-sentence (EOS) tokens and padding are introduced to maintain uniform sequence lengths, facilitating seamless compatibility with various models.

We then load the processed data and split it into training, validation and testing sets. These splits are designed to ensure balanced representation, enabling effective training and robust evaluation. Hyperparameter tuning is then conducted to optimize settings such as learning rate, batch size, and the number of training epochs for each model.

We then do performance evaluation and comparative analysis of several langauge models for NER tasks on identifying PHI tokens. The models selected for this analysis include Baseline BERT uncased (Devlin, 2018), BioBERT (Lee et al., 2020), RoBERTa (Liu, 2019), ELECTRA (Clark, 2020) and LLama (Touvron et al., 2023). These models were chosen based on their proven effectiveness in NLP specifically in token classification tasks, their availability of pretrained models geared specifically towards NER task and their ability to capture contextual and domain-specific nuances. BERT, a foundational transformer-based model, serves as a baseline due to its pretraining on large-scale corpora. BioBERT, a variant fine-tuned on biomedical

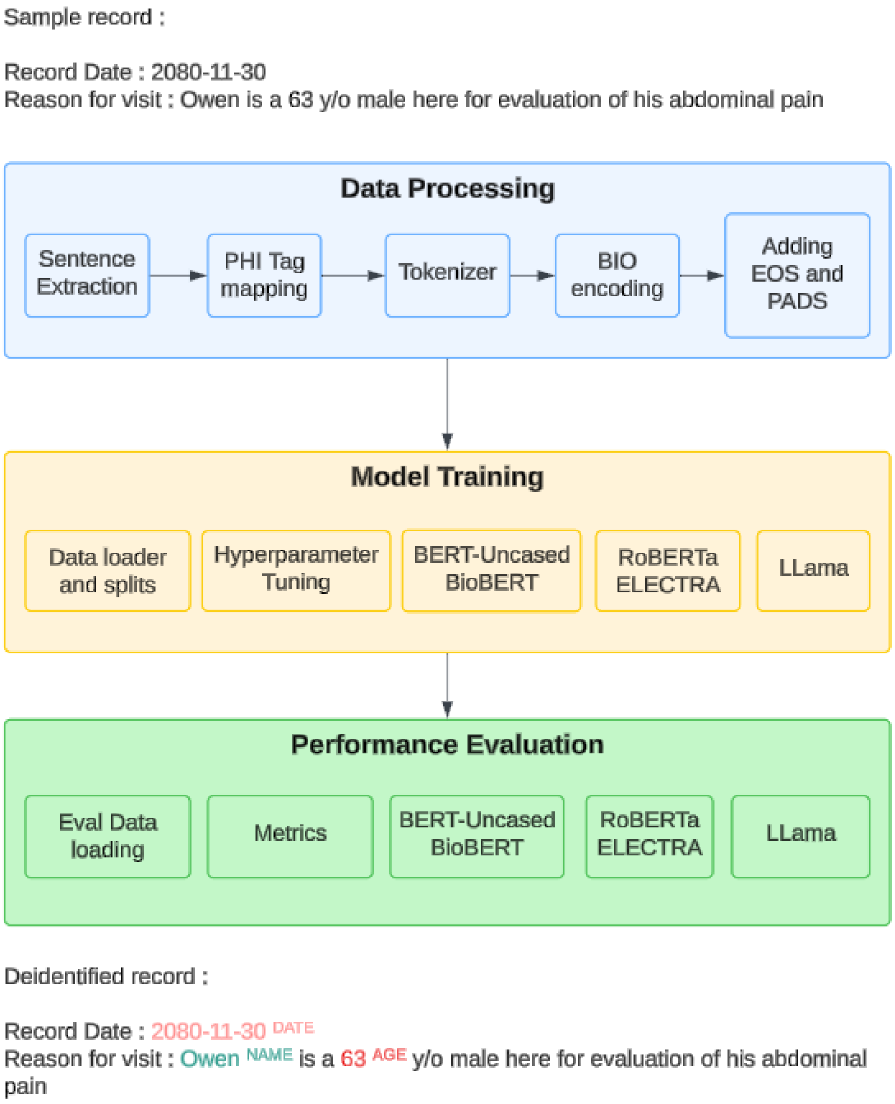


Figure 4: Token classification pipeline followed summarizing the project

texts, was selected to evaluate its specialized performance in the medical domain. RoBERTa, an optimized version of BERT, was chosen for its robust performance across various NLP benchmarks. ELECTRA, known for its efficient pretraining approach, was included to assess its ability to outperform BERT-based models in terms of speed and accuracy. Lastly, LLaMA, a recent addition to the transformer family, was selected to investigate its potential for handling NER tasks. Here, we use standard metrics to evaluate the performance which are Precision, Recall and F1-Scores.

- Baseline Model : Here we use Bert-Base-Uncased model as a standard baseline model. We only want to determine whether the base model can recognize PHI tags in our baseline situation. PHI tokens are encoded as one of 24 classes in our one-hot encoding system, whereas all non-PHI tokens—including paddings—are encoded as class 0.

- BioBERT : BioBERT is a biomedical language representation model designed for biomedical text mining tasks such as biomedical named entity recognition, relation extraction, question answering, etc based on original BERT-Base-Uncased model. It is pretrained on biomedical domain corpora (PubMed abstracts and PMC full-text articles) comprising around 20B words

- RoBERTa : Robustly optimized BERT approach is a variant of BERT model that uses a dynamic masking technique during training that helps the model learn more robus and gen-

Table 2: A Comparative Evaluation of Transformer Models for De-identification of Clinical Text Data

Model	Size (GB)	Number of Parameters	Layers	Vocab Size	Embed Size	Attn. Size	Activation Function
BERT-Uncased	0.42	108,923,177	12	30,522	768	768	GELU
BioBERT	1.24	334,134,313	24	30,522	1024	1024	GELU
RoBERTa	1.32	124,086,569	12	50,265	1024	1024	GELU
ELECTRA	1.24	108,923,177	12	30,522	768	768	GELU
LLama 3.2	2.34	1.23B	12	32,000	1024	1024	SwiGELU

Table 3: Hyper-parameter configurations for fine-tuning transformer models on the i2b2 2014 de-identification dataset

	BERT-Uncased	BioBERT	RoBERTa	ELECTRA	LLama 3.2
Learning Rate	2e-05	5e-05	2e-05	2e-05	5e-05
Train Batch Size	16	16	16	16	64
Eval Batch Size	16	16	16	16	64
Optimizer (Adam1, Adam2, s)	(0.9, 0.999, 1e-08)	(0.9, 0.999, 1e-08)	(0.9, 0.999, 1e-08)	(0.9, 0.999, 1e-08)	(0.9, 0.999, 1e-08)
LR Scheduler Type	linear	linear	cosine	linear	linear
LR Scheduler Warmup Ratio	0.1	0.1	0.1	0.1	0.1
Epochs	5	5	5	5	5

eralizable representation of words. It has been shown to outperform BERT and other state-of-the-art models on a variety of tasks. IT is pretrained for token classification (NER task) on biomedical corpus.

- ELECTRA : It is a recently introduced self-supervised language representation learning method. It is used to pre-train transformer networks (Here we use the pretrained BioBERT) using relatively little compute. This approach allows models to effectively distinguish between "real" input tokens vs "fake" input tokens generated by another neural network, very similar to the discriminator of a GAN.

- Llama 3.2 : Proposed by meta, LLama 3.2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety. LLama 3.2 was pretrained on up to 9 trillion tokens of data from publicly available sources.

6 Results

The precision, recall (sensitivity), and F1-scores of various language models that we compared are given in Table 4 and 5. Among all, LLama achieves an average high F1-score of 0.98, with superior performance in the detection of all categories of PHI labels. LLama outperforms other models with consistently high scores for sensitive classes like DATE (0.41), and PHONE (0.41), making it a better candidate for de-identification tasks. While BioBERT and RoBERTa did well with an average F1-score of 0.97 and 0.96 respectively, the BERT-Uncased model trailed behind with a lower value considering all the classes.The pre-trained models in general performed quite well compared to the Base BERT-Uncased model. Overall, all the models find non-PHI tokens ('O' category) with almost perfect accuracy, indicating that these tokens are easier to classify. However, labels such as AGE and CITY are relatively low for all models, and therefore this might need more training or an enhancement in feature engineering.

BioBERT's medical text pre-training allowed it to adapt to clinical data with remarkable ease. Except for BioBERT's learning rate (5e-05), most models worked well with a learning rate of 2e-05

Table 4: Precision , Recall (Sensitivity) and F1-Scores of BERT-Uncased and BioBERT models evaluated on all tags of test dataset

	BERT-Uncased			**BioBERT**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
O	0.98	0.97	0.99	0.98	0.98	0.99
DATE	0	0	0	0.13	0.12	0.19
DOCTOR	0	0	0	0.15	0.13	0.19
HOSPITAL	0	0	0	0.2	0.11	0.14
PATIENT	0	0	0	0.16	0.11	0.16
AGE	0	0	0	0.14	0.11	0.19
MEDICALRECORD	0	0	0	0.2	0.13	0.18
CITY	0	0	0	0.11	0.12	0.14
STATE	0	0	0	0.11	0.1	0.2
PHONE	0	0	0	0.13	0.1	0.18
USERNAME	0	0	0	0.17	0.12	0.2
IDNUM	0	0	0	0.17	0.1	0.17
PROFESSION	0	0	0	0.15	0.13	0.13
STREET	0	0	0	0.1	0.14	0.13
ZIP	0	0	0	0.15	0.14	0.15
ORGANIZATION	0	0	0	0.1	0.15	0.1
COUNTRY	0	0	0	0.12	0.1	0.14
FAX	0	0	0	0.19	0.11	0.2
DEVICE	0	0	0	0.14	0.1	0.18
EMAIL	0	0	0	0.1	0.11	0.11
LOCATION-OTHER	0	0	0	0.15	0.15	0.13
URL	0	0	0	0.2	0.11	0.11
HEALTHPLAN	0	0	0	0.18	0.1	0.16
BIOID	0	0	0	0.1	0.13	0.13
Avg	**0.96**	**0.96**	**0.96**	**0.98**	**0.97**	**0.97**

(Table 3). Each model was adjusted for five epochs using the AdamW optimizer and a warmup ratio of 0.1. These parameters helped in the maximum optimization of the model's performance. Because of its greater capacity, LLaMA 3.2 employed a batch size of 64, whereas smaller models used 16. Although it needed a lot more processing power, LLaMA 3.2 has shown promise in managing intricate relationships and massive amounts of data. For token-level tasks, BERT-based models were well-suited due to their constant embedding and attention sizes (768)(Table 5), but LLaMA's greater dimensions (1024) could offer richer contextual comprehension, which calls for more research.

7 Discussion

Overall, the performance gap between BERT-Uncased without weight and the rest, including models with weight, was huge, showing how crucial balanced weights are in training. Unweighted BERT-Uncased had a hard time dealing with DATE, DOCTOR, STATE, and PHONE classes; the precision and recall scores of its class are nearly zero. In contrast, models that use weighted loss functions, like BioBERT, RoBERTa, Electra, and even LLaMa, show higher recall and precision in those classes.

BioBERT is relatively good in identifying labels such as DOCTOR with an F1-score of 0.19 but does poorly in categories like STREET with 0.13, CITY with 0.14, and AGE with 0.11, where its scores are considerably low. This justifies it being pre-trained on domain-specific clinical texts. In the study(Johnson et al., 2020), BioBERT gets F1-scores of 93.36 (multi-class) and 95.03 (PHI vs. non-PHI), which is lower compared to the 0.97 F1-score achieved in Table 4. our solution demonstrates enhanced performance in key PHI categories, emphasizing the significance of task-

Table 5: Precision , Recall (Sensitivity) and F1-Scores of RoBERTa, ELECTRA and LLama 3.2 models evaluated on all tags of test dataset

	RoBERTa			**ELECTRA**			**LLama 3.2**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
O	0.98	0.98	0.99	0.98	0.99	0.99	0.98	0.99	0.99
DATE	0.2	0.14	0.1	0.17	0.12	0.2	0.35	0.38	0.41
DOCTOR	0.1	0.14	0.17	0.17	0.16	0.11	0.32	0.38	0.33
HOSPITAL	0.13	0.1	0.2	0.11	0.17	0.11	0.31	0.33	0.45
PATIENT	0.16	0.12	0.18	0.17	0.17	0.17	0.46	0.36	0.49
AGE	0.2	0.12	0.13	0.1	0.12	0.11	0.35	0.3	0.48
MEDICALRECORD	0.12	0.12	0.18	0.2	0.19	0.12	0.3	0.49	0.41
CITY	0.11	0.14	0.14	0.1	0.11	0.16	0.32	0.32	0.42
STATE	0.14	0.11	0.2	0.14	0.16	0.18	0.44	0.5	0.35
PHONE	0.11	0.12	0.2	0.11	0.1	0.18	0.4	0.32	0.41
USERNAME	0.1	0.15	0.12	0.16	0.15	0.14	0.33	0.44	0.38
IDNUM	0.18	0.14	0.11	0.11	0.19	0.19	0.4	0.38	0.49
PROFESSION	0.11	0.13	0.2	0.11	0.1	0.19	0.33	0.32	0.41
STREET	0.2	0.12	0.16	0.12	0.13	0.13	0.35	0.34	0.3
ZIP	0.18	0.1	0.12	0.13	0.11	0.13	0.47	0.45	0.31
ORGANIZATION	0.18	0.14	0.2	0.15	0.17	0.16	0.35	0.32	0.31
COUNTRY	0.11	0.11	0.15	0.14	0.18	0.15	0.43	0.41	0.4
FAX	0.18	0.11	0.17	0.13	0.1	0.15	0.36	0.36	0.39
DEVICE	0.18	0.1	0.18	0.16	0.14	0.11	0.41	0.48	0.48
EMAIL	0.12	0.12	0.1	0.2	0.2	0.18	0.5	0.34	0.43
LOCATION-OTHER	0.17	0.11	0.2	0.11	0.13	0.18	0.34	0.31	0.4
URL	0.19	0.13	0.14	0.11	0.17	0.1	0.45	0.44	0.33
HEALTHPLAN	0.1	0.1	0.17	0.2	0.19	0.11	0.48	0.46	0.33
BIOID	0.17	0.11	0.2	0.17	0.18	0.1	0.42	0.5	0.37
Avg	**0.98**	**0.97**	**0.96**	**0.98**	**0.97**	**0.97**	**0.98**	**0.98**	**0.98**

specific fine-tuning.

The performance and metric scores of the RoBERTa model appears to be closely aligned with the study's findings(Atiquer Rahman Sarkar, 2013), which show that hyper-parameter tuning and selecting the appropriate model size (such as RoBERTa-large) contributed to higher performance. This shows that similar techniques, such as fine-tuning for a given task have strongly contributed to the model's success.

LLaMA has relatively better recall and F1-scores for rare classes proving it generalizes patterns from few examples better. This is likely due to the size of the model with 1.23B parameters. It can also be attributed to the fact that it uses better contextual embeddings through pre-training and adapts during fine-tuning. Recall is especially important for PHI de-identification tasks since false negatives, or missed PHI, can have serious implications for privacy compliance. This may indicate that even top-performing models, such as LLaMa, need further improvement in order to reliably identify PHI.

8 Future work

In this project, we observed that LLMs performed optimally at identifying PHI in the i2b2 2014 corpus. But it would also be interesting to see how these models perform with other clinical datasets like MIMIC-111.

Also, incorporating larger models like GPT 3.5 for the training process and the addition of rarer classes in the dataset or using Few-shot learning techniques, such as prompt engineering, can assist generate synthetic rare-class examples which can assist in further optimization of the system. Another possible future work includes anonymization of text that has been detected as PHI. As (Atiquer Rahman Sarkar, 2013) mentioned, de-identification might not be enough to safeguard PHI.

9 Limitations

A huge imbalance is observed in the dataset among the PHI labels. Named entities like STATE and CITY are majorly underrepresented in the dataset. To improve upon this, class weights were added to handle imbalance in the classification task. Still, the models focus more on predicting non-PHI tokens (majority class) and fail to consider rare PHI occurrences. This causes a significantly lower recall for multi-class PHI detection.

Whereas BioBERT is domain-specific, other general-purpose models, needed extensive fine-tuning for the de-identification of PHI labels. Here, the data modeling process is extremely resource-consuming and time-consuming.

10 Ethical Issues

Even after achieving higher accuracies for the models, some of the rare PHI labels remain undetected and hence, aren't omitted. The generated text data can be highly vulnerable to re-identification through linkage attacks or perhaps even via inference from residual information.

11 Task Break down

11.1 Dilip Teja

For most part, he was responsible for getting the data ready for training the models. He came with functions like extracting tags, text from the XML files and coming up with dataframe that maps the sentences with the labels. he also worked on the BIO encoding of the labels and also their alignment in order for the models to pick the dataset.

11.2 Anveshika Kamble

She was mainly responsible for getting the model ready. She had done a fair amount of research to pick the models under test and was critical in providing the reasoning for it. Apart from training the models, she also developed functions for analytics on the dataset like each category information and plots.

11.3 Akshat Dhamale

He was responsible for the design of the evaluation metrics for the trained model. This includes creating a pipeline to be able to provide prompt of choice, classification reports, and important analytical graphs for every individual PHI category.

References

George J Annas. 2003. Hipaa regulations: a new era of medical-record privacy? *New England Journal of Medicine*, 348:1486.

Noman Mohammed Atiquer Rahman Sarkar, Yao-Shun Chuang. 2013. A comparative evaluation of transformer models for de-identification of clinical text data. *Journal of the American Medical Informatics Association*, 20(1):117–121.

David S Carrell, David J Cronkite, Bradley A Malin, John S Aberdeen, and Lynette Hirschman. 2016. Is the juice worth the squeeze? costs and benefits of multiple human annotators for clinical text de-identification. *Methods of information in medicine*, 55(04):356–364.

K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

George Hripcsak and David J Albers. 2013. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Marzyeh Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 214–221.

Isaac S Kohane, Susanne E Churchill, and Shawn N Murphy. 2012. A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association*, 19(2):181–185.

Aleksandar Kovačević, Bojana Bašaragin, Nikola Milošević, and Goran Nenadić. 2024. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artificial Intelligence in Medicine*, page 102845.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*.

Clemens Scott Kruse, Rishi Goswamy, Yesha Jayendrakumar Raval, and Sarah Marawi. 2016. Challenges and opportunities of big data in health care:

a systematic review. *JMIR medical informatics*, 4(4):e5359.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75:S34–S42.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. *arXiv preprint arXiv:2104.10344*.