Generation and De-Identification of Indian Clinical Discharge Summaries using LLMs

Sanjeet Singh[‡*] **Shreya Gupta**[†*] **Niralee Gupta**[†]
Naimish Sharma[†] **Lokesh Srivastava**[†] **Vibhu Agarwal**[†]
Ashutosh Modi[‡]
[‡]Indian Institute of Technology Kanpur (IIT Kanpur) [†]Miimansa
{sanjeet, ashutoshm}@cse.iitk.ac.in
{shreya.gupta,niralee.gupta,naimish.sharma}@miimansa.com
{lokesh.srivastava,vibhu}@miimansa.com

Abstract

The consequences of a healthcare data breach can be devastating for the patients, providers, and payers. The average financial impact of a data breach in recent months has been estimated to be close to USD 10 million. This is especially significant for healthcare organizations in India that are managing rapid digitization while still establishing data governance procedures that align with the letter and spirit of the law. Computer-based systems for de-identification of personal information are vulnerable to data drift, often rendering them ineffective in cross-institution settings. Therefore, a rigorous assessment of existing de-identification against local health datasets is imperative to support the safe adoption of digital health initiatives in India. Using a small set of de-identified patient discharge summaries provided by an Indian healthcare institution, in this paper, we report the nominal performance of de-identification algorithms (based on language models) trained on publicly available non-Indian datasets, pointing towards a lack of cross-institutional generalization. Similarly, experimentation with off-the-shelf de-identification systems reveals potential risks associated with the approach. To overcome data scarcity, we explore generating synthetic clinical reports (using publicly available and Indian summaries) by performing in-context learning over Large Language Models (LLMs). Our experiments demonstrate the use of generated reports as an effective strategy for creating high-performing de-identification systems with good generalization capabilities.

1 Introduction

Over 330 million patient records in India have already been linked with a unique central ID (PIB Press Release). To put this in perspective, the number roughly equals the total population of the United States. Several federal initiatives aimed at establishing standards for medical information exchange, adoption of controlled terminologies, and promoting open architecture-based systems for the management of patient records have seen a steady rise in the adoption of electronic health records within Indian healthcare institutions (Ministry of Health and Family Welfare (MoHFW), India; Srivastava, 2016). This data represents an under-utilized resource that has profound implications for informing public policy, medical research and patient care. At the same time, it also poses some serious challenges. The risks of revealing patient identity even from data that has been anonymized are well known (Sweeney, 2013). Privacy regulations such as GDPR 2016 (European Parliament and Council of the European Union) and the HIPAA Privacy Rule 2003 (U.S. Department of Health and Human Services (HHS)) lay down heavy penalties on non compliance with data safety protocols. A robust data de-identification pipeline is vital if we aim to unlock insights from these electronic patient histories.

Natural Language Processing (NLP) methods for de-identification are known to perform significantly better than manual de-identification (Douglass et al., 2004). However, these have been studied mostly in the single-institution setting. There are limited studies that evaluate de-identification performance of these methods across institutions (Yang et al., 2019). These suggest that NLP methods for de-identification perform poorly when evaluated on data from a different institution compared to the one that contributed the training data. This is especially significant in the context of patient data originating within Indian healthcare institutions. To the best of our knowledge, studies evaluating the performance of NLP based de-identification systems on patient data from Indian healthcare institutions have not yet been carried out. One reason for this might be that until recently there was no regulatory framework for accessing patient data for

[*]Equal Contribution

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 342–362

research. The Indian Digital Personal Data Protection Act 2023 (DPDPA) (Ministry of Electronics and Information Technology (MeitY), India) is a landmark legislation that came into effect in September 2023 and covers all organizations that process the personal data of individuals in India. Similar to GDPR 2016, the DPDPA defines responsibilities for organizations that collect, store, and process data from patients in India and holds them legally accountable for safeguarding patient privacy. The DPDPA also highlights the need for a data de-identification solution that has been validated on patient data from Indian healthcare institutions.

The present study takes a step towards answering this imminent need. Using a dataset of fully de-identified 99 discharge summaries obtained under Institutional Review Board (IRB) approval from the Sanjay Gandhi Post Graduate Institute of Medical Sciences (SGPGIMS), Lucknow, India, the study evaluates language models (LMs) for the task of de-identification. Furthermore, commercially available de-identification solutions are also evaluated. Hereafter, we refer to this dataset as the Indian Clinical Discharge Summaries (ICDS$_R$, subscript R refers to real) dataset. Given the paucity of clinical data, the study also evaluates Large Language Models (LLMs) on the task of generating synthetic clinical texts for training de-identification models. Critically, the study highlights the existence of several personal health information (PHI) elements in the ICDS$_R$ dataset that are unique to the language use and cultural practices in India. It is unlikely that the existing de-identification solutions have been trained to recognize these unique PHI elements, and therefore, their detection may be unreliable. In a nutshell, we make the following contributions:

- We introduce a new dataset (Indian Clinical Discharge Summaries (ICDS$_R$)) obtained from an Indian hospital and evaluate the performance of PI-RoBERTa model (PI-RoBERTa) (fine-tuned on non-Indian clinical summaries) on ICDS$_R$ for the task of De-Identification. Our experiments show poor cross-institutional performance. Experiments with existing commercial off-the-shelf clinical de-identification systems show similar trends.
- To overcome the paucity of Indian clinical data, we generate synthetic summaries using LLMs (Gemini (Team et al., 2023), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), and Llama3 (Touvron et al., 2023)) via In-Context Learning (ICL). Further, the synthetic summaries are used to train PI-RoBERTa for de-identification on ICDS$_R$. Results show significant improvement in the performance of the de-identification system.
- We release the model code and experiments via GitHub: `https://github.com/Exploration-Lab/llm-for-clinical-report-generation-deidentification`

2 Related Work

Automatic data de-identification methods for biomedical texts have focused on leveraging machine learning techniques to ensure privacy while maintaining data utility. Named Entity Recognition (NER) systems have been tailored to identify and anonymize personal health information/personal identifiable information (PHI/PII) from clinical narratives. Earlier work explored Support Vector Machines (SVMs) for identifying PHI (Neamatullah et al., 2008). Researchers have also explored deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Dernoncourt et al., 2017), which have shown superior performance over the conventional approach.

In recent years, there has been a growing interest in the application of transformer-based models like BERT (Devlin et al., 2018) for the clinical NER and de-identification task (Chaudhry et al., 2022; Alsentzer et al., 2019). LLMs have also been explored for various clinical tasks such as clinical NLI (Mandal and Modi, 2024). Hybrid approaches that combine rule-based and machine learning methods have also been developed to enhance the robustness of de-identification systems (Meystre et al., 2010). A study by Yang et al. (2019) used a hybrid model combining Long Short-Term Memory (LSTM) networks with Conditional Random Fields (CRFs) for the de-identification of clinical notes. It demonstrated the effectiveness of integrating local resources and diverse word embeddings, and achieved high F1 scores across various de-identification tasks. Furthermore, El Azzouzi et al. (2023) de-identified French electronic health records using distant supervision and deep learning techniques. The study utilized models like Bi-LSTM+CRF and enhanced them with contextualized word embeddings. It achieved remarkable accuracy in removing identifiable information while maintaining data utility. These innovations underscore the continuous improvement

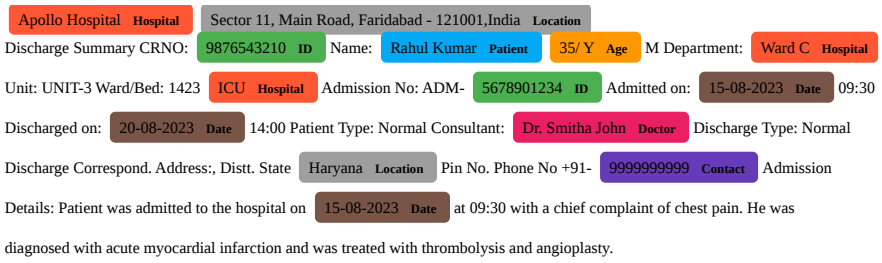


Figure 1: A sample of annotated text from Discharge Summary

and adaptation of de-identification methods to address the evolving challenges in data privacy. With remarkable progress made in generative AI techniques, researchers have started exploring generating synthetic clinical data. For example, medGAN (Choi et al., 2017) has been proposed to generate high-dimensional discrete variables such as patient records. It shows that it can produce realistic EHR data that preserves the statistical properties of the original dataset. Researchers have also explored differential privacy techniques in conjunction with Generative Adversarial Networks (GANs) to ensure that the synthetic data does not allow re-identification of individuals. There is also ongoing research into hybrid models that combine rule-based and machine learning techniques to generate data that not only looks realistic but also adheres to known clinical correlations and constraints (Isasa et al., 2024; Goncalves et al., 2020). Such approaches ensure that the synthetic data is both safe and scientifically valid for use in biomedical modeling simulations. The trend highlights the potential of synthetic data to address privacy and data availability challenges in biomedical research. In this paper, we explore LLMs for generating synthetic clinical reports that closely resemble reports in ICDS$_R$, thus capturing the underlying data generation processes.

3 Clinical Discharge Summaries Datasets

n2c2: We make use of the 2006 and 2014 n2c2 datasets (Özlem Uzuner et al., 2008; Stubbs et al., 2015). The 2006 challenge involved the development of automated methods to de-identify discharge summaries from patient medical records (Özlem Uzuner et al., 2008). The total number of summaries in the n2c2-2006 dataset are 888, split between training and test sets. The 2014 challenge comprised of two tasks: de-identification and heart disease risk factor identification (Stubbs et al., 2015). For the de-identification task, the dataset included a variety of clinical documents such as progress notes, discharge summaries, and other narrative texts that typically contain detailed patient information.

Indian Clinical Discharge Summaries (ICDS$_R$): We obtained fully de-identified 99 discharge summaries obtained under Institutional Review Board (IRB) approval from the Sanjay Gandhi Post Graduate Institute of Medical Sciences (SGPGIMS), Lucknow, India. All discharge summaries in the Indian Clinical Corpus were manually annotated for de-identified entities by human annotators using Doccanno (Nakayama et al., 2018), a data annotation tool. Each document was annotated by one annotator. The annotators had previous experience in clinical text annotations. Following established practice, we used the BIO scheme (Ramshaw and Marcus, 1999) for annotating named entities. Our PHI labels were defined by augmenting the PHI entities defined in the HIPAA Privacy Rule 2003 along with adaptation to Indian clinical texts. After annotation, we obtained 26 PHI unique entities in the ICDS$_R$ dataset. Subsequently, due to privacy concerns, PHI elements were replaced with fake values through an automatic replacement tool developed using the Python library *Faker* (Faraglia and Other Contributors, 2010) (example in Fig. 1). Repeated occurrences of an entity within a note were tracked for consistent replacements. Moreover, settings such as date/time offsets were parameterized via a configurable file. The tool provides a scalable solution for de-identifying medical datasets while ensuring secure data access. Table 1 provides statistics of the datasets.

4 Generated Discharge Summaries Datasets

Initial experimentation showed over-fitting in models on the ICDS$_R$ data due to its small size (69, 10, 20 summaries for train, val, and test sets, respectively). Consequently, we generated synthetic summaries to augment ICDS$_R$ data. Synthetic patient data is being used increasingly for a variety

Statistics	Training dataset					Test set		
	n2c2-2006	n2c2-2014	$ICDS_R$	$ICDS^g_G$	$ICDS^l_G$	n2c2-2006	n2c2-2014	$ICDS_R$
# Summaries	668	790	79	1596	1043	220	514	20
# Unique Tokens	29218	55907	13542	56780	25184	15231	41066	6106
Max Length	3023	2984	9494	4256	2590	2687	2474	8511
Min Length	13	74	97	100	109	15	99	270
Avg. Summary Length	581.71	618.86	1005.94	373.80	392.34	748.22	615.19	1343.40
Original Tag Set	9	24	26	34	106	9	21	24
Mapped Tag set	9	9	9	9	9	9	9	9

Table 1: Statistics of various datasets

of in-silico biomedical experiments in addition to training data augmentation (Chen et al., 2021). Using the samples from $ICDS_R$ we generated medical discharge summaries specific to Indian patients using LLMs (Gemma, Llama-3-8B-Instruct, and Mistral-7B-Instruct-v0.1) via In-Context Learning (ICL). We experimented extensively with various prompts and discharge summaries, as explained below. Our choice of LLMs was driven by the feasibility of instantiating these models on-premise. Prompting is a key aspect of using LLMs. As described below, we experimented with various prompt designs.

Discharge Summaries Generation using the n2c2-2006 dataset: Since the n2c2-2006 discharge summaries are publicly accessible, we generated synthetic discharge summaries based on these along with PHI annotations using Gemini-pro-1.0. We arrived at a functional prompt by iteratively tuning and inspecting the synthesized outputs for overall length, presence of key subsections, and correct PHI annotation. While tuning our prompts, we did not check for the medical validity of the discharge summaries (see App. Table 12). The prompt also contained an original n2c2-2006 summary as an exemplar. This way, we generated five patient discharge summaries for each original discharge summary in the n2c2-2006 dataset and a total of 3000 discharge summaries. The generated summaries were manually reviewed, and the ones containing gibberish text and missing or incorrect annotations were filtered out, resulting in 1596 synthetic discharge summaries with PHI annotations. Hereinafter, we refer to this dataset as $ICDS^g_G$.

Discharge Summaries Generated using the $ICDS_R$ dataset: The $ICDS_R$ dataset is accessible only under the Institutional Review Board's approval, and therefore, LLMs that can be inferred only via public API endpoints cannot be used to process these. Consequently, we generated synthetic discharge summaries for the $ICDS_R$ dataset only with LLMs that could be instantiated within our secure compute infrastructure (Llama-3-8B-Instruct, Gemma and Mistral-7B-Instruct-v0.1, respectively). We evaluated various LLM and prompt combinations to converge on Llama-3-8B-Instruct (see App. Table 12 for the prompt). To evaluate the performance of model-prompt combinations, we calculated two metrics: BERT F1-Score and the average length of summaries (in words). The BERT F1-Score was calculated on a sample of synthetic annotated discharge summaries (target) and the 99 original $ICDS_R$ discharge summaries (see Table 2). The BERT F1-Score of Meta-Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.1 models with prompt B surpass other model-prompt combinations. We selected the Meta-Llama-3-8B-Instruct model for synthetic discharge summary generation and PHI annotation since, in addition to a high BERT F1-score, the generated summaries are, on average, longer. The $ICDS_R$ dataset was split so that 79 summaries were used in the prompt to generate synthetic summaries while the remaining 20 were reserved for the test set. The temperature parameter of Meta-Llama-3-8B-Instruct was set to 0.9. Around 25 summaries were generated for each of the 79 $ICDS_R$ discharge summaries by embedding these one at a time as an exemplar in the prompt. In total, 1831 discharge summaries, which already had PHI annotations, were generated, yielding 1043 generated discharge summaries after manual review and filtration. Hereinafter, we refer to this dataset as $ICDS^l_G$. Further, we asked two annotators to annotate 50 generated summaries (after removing the PHI tags) with PHI tags. The Cohen's kappa coefficient (Warrens, 2015), the measure of inter-annotator agreement, was 0.921, showing a high agreement.

Evaluation of the Quality of the Generated Summaries: We assessed the face validity of the gen-

Prompt Id	Model Used	BERT F1-Score	Avg. Summary Length (words)
B	meta-llama/Meta-Llama-3-8B-Instruct	0.491	564
B	mistralai/Mistral-7B-Instruct-v0.1	0.493	400
C	meta-llama/Meta-Llama-3-8B-Instruct	0.486	503
C	mistralai/Mistral-7B-Instruct-v0.1	0.468	267
C	google/gemma-1.1-7b-it	0.478	268

Table 2: Comparison of model-prompt combinations

Training Set	Test Set		
	n2c2-2006	n2c2-2014	ICDS_R
n2c2-2006	✓	✓	✓
n2c2-2014	✓	✓	✓
n2c2-2006+ n2c2-2014	✓	✓	✓
ICDS^g_G	✓	✓	✓
ICDS^l_G	✓	✓	✓
ICDS^g_G + ICDS^l_G	✓	✓	✓
ICDS^g_G + ICDS^l_G + n2c2-2014	✓	✓	✓
ICDS^g_G + ICDS^l_G + n2c2-2014	✓	✓	✓

Table 3: Experiments Matrix

erated summaries by asking physicians to review a convenience sample of 30 real and 30 synthetic discharge summaries with the real/synthetic labels suppressed. The 60 discharge summaries were shuffled and uploaded to a secure, online review tool accessible only to the reviewers (physicians). The reviewers were asked to review each summary and then assign a single label (real or synthetic) to each based on their experience. The review results were compiled, and the precision, recall, and F1 scores were computed for each physician along with Cohen's Kappa to assess agreement between the two physicians (details in §7).

As can be observed in Table 1, for the purpose of uniformity and modeling, we mapped PHI entities in each of the dataset to 9 tags (corresponding to 8 unique entities + 1 `OTHERS`). App. Table 15 provides details of tag mapping where the PHI entities are mapped with to their superset and all non-PHI entities are mapped to `OTHERS` Tag.

5 De-Identification Task

De-Identification Task: De-Identification is conceptually similar to a Named Entity Recognition task. Both ICDS^g_G and ICDS^l_G were pre-processed and converted into BIO format as is customary in Named Entity Recognition development (also see App. Fig. 4). Formally, given some text, $S = (w_1, w_2, w_3, ..., w_n)$ containing n words, de-identification requires labeling each of the word w_i with a tag t_k coming from a NER tagset $t_1, t_2, ..., t_T$. Subsequently, the labeled entities can be redacted or replaced with fake values for privacy protection.

De-Identification Model: We fine-tuned several different NER models, including ghadeermobasher/BCHEM4-Modified-BioBERT-v1 (BioBERT) and Clinical-AI-Apollo/Medical-NER (Clinical AI Apollo). In each case, we used a training partition of the data to train and a validation partition for evaluation. However, the Clinical NER models did not perform well since they are designed to label medical entities such as disease, drugs, procedures, and devices (see App. D). RoBERTa-NER-Personal-Info model (PI-RoBERTa) showed good performance on n2c2-2006 and n2c2-2014 datasets. The architecture for PI-RoBERTa is shown in App. Fig. 13. PI-RoBERTa is a 24-layered transformer model that predicts a label for each token.

Attribute	Dataset	
	Real	Generated
Counts	3158684	5022667
Length (words)	560753	721886
Mean ± SE	4.64 ± 0.004	5.93 ± 0.005
Median	4.0	5.0
Min	1	1
Max	50	89
Jaccard Distance	0.83	
BERTScore (F1)	0.64	
BERTScore (Precision)	0.65	
BERTScore (Recall)	0.63	

Table 4: Comparison of n2c2-2006 and ICDS^g_G Dataset

Attribute	Dataset	
	Real	Generated
Counts	636805	4789863
Length (words)	102604	508244
Mean ± SE	5.21 ± 0.01	7.77 ± 0.01
Median	4.0	5.0
Min	1	1
Max	72	472
Jaccard Distance	0.80	
BERTScore (F1)	0.58	
BERTScore (Precision)	0.60	
BERTScore (Recall)	0.56	

Table 5: Comparison of ICDS_R and ICDS^l_G Dataset

6 Model Training Experiments

Initial experiments with ICDS_R using a 69-10-20 (train-val-test) split resulted in overfitting given that ICDS_R is small, containing only 99 discharge summaries. We also experimented with training the model on n2c2-2006 and n2c2-2014 datasets and testing on ICDS_R to check for cross-institutional generalization. We experimented with several combinations of real and synthetic datasets and evaluated on the test set of n2c2-2006, n2c2-2014, and ICDS_R. Table 3 shows the experiments matrix, in total we evaluated 24 different combinations. For all the experiments, we reserved 20 summaries of ICDS_R for testing. Note that these summaries were also not used for generation. For each experiment, PI-RoBERTa was fine-tuned on each training set as

Training Data	n2c2-2006			n2c2-2014			n2c2-2006 + n2c2-2014		
Testing Data	n2c2-2006	n2c2-2014	ICDS$_R$	n2c2-2006	n2c2-2014	ICDS$_R$	n2c2-2006	n2c2-2014	ICDS$_R$
CONTACT	0.98	0.66	0.18	0.73	0.95	0.20	0.96	0.93	0.24
PATIENT	0.95	0.65	0.81	0.82	0.98	0.85	0.91	0.96	0.77
DOCTOR	0.95	0.89	0.64	0.93	0.98	0.76	0.97	0.98	0.54
ID	0.99	0.55	0.64	0.96	0.97	0.65	1.00	0.96	0.93
DATE	0.98	0.43	0.16	0.70	0.99	0.97	0.97	0,98	0.97
LOCATION	0.89	0.80	0.71	0.78	0.95	0.80	0.81	0.94	0.75
HOSPITAL	0.94	0.79	0.34	0.87	0.94	0.36	0.94	0.93	0.40
AGE	0.80	0.00	0.00	0.02	0.99	0.48	0.12	0.94	0.53
Micro Avg	0.96	0.66	0.41	0.81	0.98	0.80	0.96	0.97	0.80
Macro Avg	0.93	0.60	0.43	0.72	0.97	0.63	0.83	0.95	0.64
Weighted Avg	0.96	0.61	0.31	0.84	0.98	0.78	0.96	0.97	0.78

Table 6: F1 scores for PHI entities with overall micro Avg F1 , macro Avg F1 , Weighted Avg F1

Training Data	ICDS^g_G			ICDS^l_G			ICDS^g_G+ ICDS^l_G		
Testing Data	n2c2-2006	n2c2-2014	ICDS$_R$	n2c2-2006	n2c2-2014	ICDS$_R$	n2c2-2006	n2c2-2014	ICDS$_R$
CONTACT	0.80	0.47	0.11	0.55	0.38	0.96	0.93	0.67	0.98
PATIENT	0.74	0.56	0.68	0.05	0.32	0.95	0.83	0.60	0.90
DOCTOR	0.86	0.78	0.88	0.35	0.71	0.98	0.86	0.76	0.98
ID	0.87	0.58	0.51	0.81	0.61	1.00	0.93	0.63	0.98
DATE	0.87	0.90	0.88	0.70	0.84	0.99	0.90	0.88	0.99
LOCATION	0.71	0.78	0.34	0.50	0.66	0.97	0.75	0.81	0.96
HOSPITAL	0.87	0.72	0.31	0.42	0.51	0.97	0.88	0.70	0.98
AGE	0.02	0.67	0.51	0.02	0.38	0.96	0.06	0.56	0.97
Micro Avg	0.85	0.77	0.68	0.55	0.67	0.98	0.88	0.76	0.98
Macro Avg	0.72	0.68	0.53	0.42	0.55	0.97	0.77	0.70	0.97
Weighted Avg	0.86	0.77	0.69	0.52	0.66	0.98	0.88	0.77	0.98

Table 7: F1 scores for PHI entities for the PI-RoBERTa trained on generated data.

given in Table 3 and tested on each corresponding test set. Details about training are given in App. D

Comparison with Commercial De-Identification Systems: We compared the performance of these on the ICDS$_R$ test set. In particular, we evaluated AWS's (Amazon Web Services) Comprehend Medical DetectPHI (Amazon Web Services) and GCP's (Google Cloud Platform) Data Loss Protection (DLP) (Google Cloud) de-identification solutions. For comparison and evaluation, ICDS$_R$ test set was mapped to a common tag set, which includes DATE, NAME, LOCATION, AGE, ID, and CONTACT. To ensure consistency across the dataset, pre-processing steps were applied. For instance, titles such as 'Dr.' and 'Mr.' were removed from NAME entities in the ICDS$_R$ test set due to the solution's inability to recognize them. Certain tags and entities were excluded from the analysis to align with a common tag set. The LOCATION entity was standardized by merging all location-related entities (street, city, state, zip) into a single LOCATION entity. Similarly, HOSPITAL, ORGANISATION_NAME and ADDRESS entities were consistently mapped to LOCATION.

De-identification using LLMs: We further evaluated the performance of LLMs on ICDS$_R$ test set. Meta-Llama-3-8B-Instruct was instantiated within our secure compute infrastructure, and the prompt was developed for medical text de-identification using the iterative approaches described in the foregoing sections.

7 Experiments, Results and Analysis

Comparison of datasets: The total number of summaries in the n2c2-2006 dataset are 888, split between training and test sets, as shown in Table 1. The n-gram analysis of the n2c2-2006 and ICDS$_R$ datasets reveals distinct linguistic patterns reflecting their unique clinical foci. The n2c2-2006 dataset features unigrams like 'patient,' 'discharge,' and medication-related terms such as 'mg' and 'po' and bigrams like 'mg po' and 'discharge date,' highlighting a narrative centered on patient management and clinical processes as shown in App. Fig. 14. In contrast, the ICDS$_R$ dataset (as shown in App. Fig. 18) shows a marked presence of terms

Training Data	n2c2-2014+ ICDS$^{l}_{G}$+ ICDS$^{g}_{G}$			n2c2-2014+ n2c2-2006+ ICDS$^{g}_{G}$+ ICDS$^{l}_{G}$		
Testing Data	n2c2-2006	n2c2-2014	ICDS$_R$	n2c2-2006	n2c2-2014	ICDS$_R$
CONTACT	0.89	0.95	0.98	0.97	0.96	0.98
PATIENT	0.87	0.97	0.88	0.94	0.96	0.88
DOCTOR	0.95	0.97	0.98	0.98	0.98	0.99
ID	0.99	0.97	0.99	0.99	0.96	0.99
DATE	0.82	0.99	0.99	0.99	0.99	0.99
LOCATION	0.76	0.95	0.98	0.85	0.94	0.98
HOSPITAL	0.93	0.94	0.96	0.96	0.94	0.97
AGE	0.02	0.97	0.96	0.35	0.97	0.86
Micro Avg	0.88	0.97	0.97	0.97	0.97	0.97
Macro Avg	0.78	0.96	0.96	0.88	0.96	0.96
Weighted Avg	0.90	0.97	0.97	0.97	0.97	0.97

Table 8: F1 scores of PHI entities when PI-RoBERTa is fine-tuned on combination of datasets

Metric	AWS	GCP
F1 Score	0.37	0.47

Table 9: Results: AWS vs. GCP Solutions on ICDS$_R$ test set

Entity	AWS	GCP
DATE	0.39	0.56
NAME	0.57	0.52
LOCATION	0.20	0.22
AGE	0.12	0.00
ID	0.17	0.17
CONTACT	0.63	0.36

Table 10: F1 scores for Entity-Wise Comparison of AWS and GCP Solutions on ICDS$_R$ test set

such as 'pm,' 'days,' and 'mgdl,' and bigrams and trigrams like '10 days,' 'daily 10,' and 'cr x ray,' suggesting an orientation towards experimental or lab-result oriented narratives, with a particular emphasis on procedural timelines and diagnostic procedures. Hence, ICDS$_R$ focuses on a broader scope involving diagnostics and treatment monitoring.

Real versus Generated Datasets

ICDS$^{g}_{G}$ vs n2c2-2006: We analyzed the n2c2-2006 and the synthetic ICDS$^{g}_{G}$ discharge summaries in terms of summary statistics, Jaccard distance, and BERTScore (using the "dmis-lab/biobert-v1.1" model) as shown in Table 4 (Lee et al., 2020; Zhang et al., 2020). The Jaccard distance suggests a high level of lexical dissimilarity between the datasets, indicating that the synthetic dataset introduces a significant degree of variation compared to the real dataset. While indicating some differences, an F1 score of 0.6362 indicates the real and synthetic datasets have semantic overlap. An n-gram analysis of the top 10 unigrams, bigrams, and trigrams unveils the differences between the two datasets, yet also underscores their relevance to the task at hand as shown in App. Fig.14, Fig.15, Fig.16, and Fig.17. These metrics suggest that while the synthetic dataset is designed to be distinct enough to introduce useful variability, it retains a substantive semantic similarity to the real dataset. This balance is crucial when synthetic data is used for tasks such as model training, where the goal is to ensure that the model is not only trained on a diverse set of data but also remains relevant and effective when applied to real-world data. The high Jaccard distance combined with the moderate BERTScore indicates that the synthetic dataset achieves this objective by being similar enough to the real dataset to be useful, yet different enough to enhance the dataset's diversity and robustness.

ICDS$^{l}_{G}$ vs ICDS$_R$: Similar to the n2c2-2006 and ICDS$^{g}_{G}$ datasets, we analyzed the ICDS$_R$ and ICDS$^{l}_{G}$ datasets with summary statistics, Jaccard distance, and BERTScore, as shown in Table 5. The Jaccard distance suggests lexical dissimilarity implying injection of new vocabulary in the generated discharge summaries. The n-gram analysis of the top 10 unigrams, bigrams, and trigrams shows these differences (App. Fig.18, Fig.19, Fig.20, and Fig.21). The BERTScore results indicate a moderate level of semantic similarity between the real and generated datasets. The metrics suggest that the generated dataset has greater lexical variety and incorporates some additional semantic constructs.

Evaluation of The Quality of Generated Summaries: The confusion matrix on convenience sample of 60 discharge summaries evaluated by physician1 and physician2 are shown in Fig. 2 and Fig. 3 respectively. There are 10 summaries that were originally synthetic but were labeled as real by physician 1, and 19 summaries that were originally synthetic but were labeled as real by physician 2. Physician 1 is able to label summaries with higher precision and recall, i.e., higher f1-score as com-

Physician	Precision	Recall	F1-score
Physician 1	0.714	0.833	0.769
Physician 2	0.537	0.733	0.620

Table 11: Evaluation metrics of 60 discharge summaries annotated by physician 1 and physician 2

pared to physician 2 (Table 11). The Cohen's kappa coefficient, the measure of inter-annotator agreement, is 0.290 showing a fair agreement between the labels assigned by the physicians. Additionally, physician 1 observed that many of the discharge summaries that he labeled synthetic appeared to have been translated from a non-English source. Physician 2 reported some diagnosis and formatting issues among the summaries he labeled as synthetic. Additionally, physician 2 reported some errors in diagnoses, medications, and lab results, but these were not limited to the summaries he labeled as synthetic.

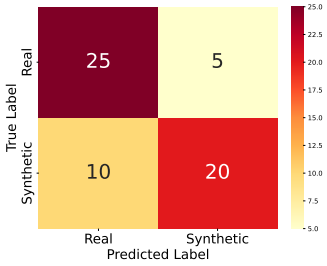


Figure 2: Confusion matrix on convenience sample (60 discharge summaries) evaluated by physician 1

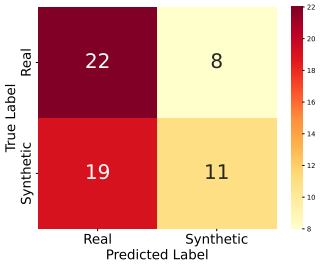


Figure 3: Confusion matrix on convenience sample (60 discharge summaries) evaluated by physician 2

Model Performance: Table 6 shows the results for intra- and inter-institutional performance. As can be observed, the inter-institutional performance of the model is very high (> 0.96 F1). However, the cross-institutional performance suffers significantly. Table 7 shows the results of training on generated datasets. The fine-tuned Model gives 68% F1 score on the ICDS$_R$ test set, 77% on the n2c2-2014 test set, and 85% on the n2c2-2006. Results on the ICDS$_R$ test set are not promising. This might have happened because ICDS^g_G was generated using n2c2-2006. Fine-tuning on ICDS^l_G dataset results in 98% F1 score on the ICDS$_R$ test set, 67% on n2c2-2014 test set, and 55% on the n2c2-2006 test set. To further improve model generalization, we experimented with combinations of datasets. Table 8 shows the training results on a combination of real and synthetic datasets. We get micro-F1 of 97% on n2c2-2014 and ICDS$_R$ test set given that we have included n2c2-2014 and ICDS^l_G datasets in training, but the performance of the model (88%) is also notable on n2c2-2006 dataset. These results indicate that fine-tuning on the combination improves cross-institutional performance.

Analysis: Our experiments indicate models have poor cross-institutional generalization. We performed several experiments with n2c2-2006, n2c2-2014, ICDS^g_G, and ICDS^l_G datasets, and their combinations. The general trend is that fine-tuned model performance degrades heavily in cross-dataset settings. At the individual entity level, the F1 score for the PATIENT entity is consistent for all the fine-tuned models. For the DOCTOR and DATE entities, the F1 scores of all the fine-tuned models are also consistent, except for when the model is trained on the n2c2-2006 dataset and tested on the n2c2-2014 dataset and ICDS$_R$ test sets. For the ID entity, all the fine-tuned models have consistent F1 scores, except for when the model is trained on ICDS^g_G, and tested on n2c2-2014 and ICDS$_R$ datasets. We noticed performance variance in the LOCATION, AGE, and CONTACT entities. This could be because the LOCATION can be any local address without a specific format. AGE is either a number like '78 Y' or a word representation of that number like 'Seventy-Eight year old'. In most cases in the datasets, these types of words or tokens are tagged as OTHERS, and they are highly prevalent. This could be why the AGE tag was incorrectly predicted as OTHERS in cross-dataset settings. The entity CONTACT includes email, IP address, phone number, landline number, etc. However, their distribution is not uniform.

Our main aim was to develop a robust model that could de-identify medical text from Indian Healthcare Institutes. This was done by fine-tuning PI-RoBERTa on ICDS^l_G where we are getting state-of-the-art performance on ICDS$_R$. Almost all the entities were correctly identified, with a few exceptions. A few PHI entities were misidentified with non-PHI entities (i.e., OTHERS) and vice versa, as can be seen in App. Fig. 26. However, the per-

centage of incorrect prediction is significantly less when considering the total support set of ICDS$_R$ test set. However, this fine-tuned model was not generalizable when we tested it on the n2c2-2006 and n2c2-2014 test sets, as seen in the Table 7. For model generalizability, we fine-tuned PI-RoBERTa on n2c2-2014 , ICDS^g_G and ICDS^l_G, tested on the n2c2-2006 test set. The results shown in Table 8 indicate that models are generalizing when we fine-tuned them on different combinations of datasets, although the F1 score for all entities is not consistent, as can be seen in App. Fig. 28a. Confusion matrix for all the experiments are shown in App. Fig. 22 Fig. 23 , Fig. 24 , Fig. 25 , Fig. 26 Fig. 27 , Fig. 28b.

Comparison with Commercial De-Identification Systems: The results obtained using AWS and GCP solutions are summarized in Table 9 and Table 10. The results clearly indicate that AWS and GCP do not perform well on ICDS$_R$ test set. This could be because systems have been trained on non-Indian specific clinical data. This underscores the importance of ensuring that de-identification caters to diverse demographics, which is essential for ensuring the efficacy and ethical deployment of these solutions.

The underperformance of commercial solutions in classifying PHI in ICDS$_R$ can be attributed to misidentification. Medical entities are mistaken as NAME/ LOCATION, while Pin-codes as ID. Names like 'Alia' and 'Adah' are not being consistently recognized as NAME by AWS and GCP. Patient IDs that start with CRNO: ########## or ADM-########## are not identified as PHI; these solutions probably aren't sure what CRNO, ADM stand for. 'B/O Kanav Viswanathan' is misidentified, where 'Kanav Viswanathan' is a name and B/O stands for Baby of but gets labeled as a LOCATION. 'Urvi Bhamini Faiyaz Kakar' is identified as Name by GCP but not by AWS. 'Wockhardt Hospitals,' hospital name was not identified as PHI. Medical terms like 'BILIRUBIN,' 'MALLOY EVELYN,' 'CR X Ray' and 'SERUM LIPASE' are misidentified as NAME when they describe medical tests. Similarly, 'CREATININE (M - JAFFE COMPENSATED)' is a medical test and 'JAFFE' is misidentified as NAME. 'Meropenem,' an antibiotic, is misidentified as NAME. Even terms like 'Ward' from room names such as 'Ward-B' occasionally get misidentified as NAME. Test results like '136/94mmHg' or 'TSH - 5.45' are misidentified as ID. Locations like 'Subramaniam Chowk' and 'Yohannan Nagar,' are also misidentified as NAME. Additionally, using GCP or AWS for PHI detection introduces variability, causing results to vary with each execution. These factors underscore the need for precision and consistency in data handling to mitigate performance issues in medical contexts.

De-identification using LLMs: We also conducted experiments of de-identifying clinical summaries using LLMs directly. A precision score of 0.55 was obtained. However, the model faced challenges in terms of recall. The recall scores were merely 0.11. We also evaluated the performance of Mistral-7B-Instruct-346v0.1 and Gemma. Surprisingly, the results obtained from these models were far inferior to those of Meta-Llama-3-8B-Instruct. Results suggest that the LLMs struggle to detect PHI in Indian medical discharge summaries.

8 Conclusion and Future Directions

In this paper, we explored the task of de-identification on Indian clinical discharge summaries. Experiments indicate a poor generalization of fine-tuned (on public datasets) models and poor performance of the off-shelf commercial systems. Experiments with LLM generated summaries look promising; the model fine-tuned on generated summaries and public datasets shows good generalization performance. Our results are based on a small test set. Using the insights from our work, we aim to set-up an active learning workflow that combines our fine-tuned model and human annotators to produce a larger test dataset on which we may evaluate overall model performance as well as by conditioning on a medical specialty. The augmented (generated summaries with original data) institution-specific dataset can be used to fine-tune NER models that have been pre-trained on PHI data cost-effectively. Achieving cross-institution portability remains a topic of active research. However, many open-source large language models can be deployed on-premise and, as described above, fine-tuned to provide an immediate and effective solution to personal data protection in Indian healthcare institutions.

9 Acknowledgements

We would like to thank Dr. Uttam Singh, Dr Prabhakar Mishra, and Dr Amit Goel for their support for this work.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, David Jindi, Tristan Naumann, and Matthew B McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Amazon Web Services. Amazon comprehend medical documentation: Analyzing protected health information (phi) [online].

BCHEM4 Modified BioBERT. ghadeermobasher/BCHEM4-Modified-BioBERT-v1 · Hugging Face — huggingface.co [online].

Mukund Chaudhry Chaudhry, Arman Kazmi, Shashank Jatav, Akhilesh Verma, Vishal Samal, Kristopher Paul, and Ashutosh Modi. 2022. Reducing inference time of biomedical ner tasks using multi-task learning. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 116–122.

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Edward Choi, Sushant Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. *arXiv preprint arXiv:1703.06490*.

Clinical AI Apollo. Clinical-AI-Apollo/Medical-NER · Hugging Face — huggingface.co [online].

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Margaret Douglass, Gari D Clifford, Andrew Reisner, George B Moody, and Roger G Mark. 2004. Computer-assisted de-identification of free text in the mimic ii database. In *Computers in Cardiology, 2004*, pages 341–344. IEEE.

Dslim bert base NER. dslim/bert-base-NER · Hugging Face — huggingface.co [online].

Mohamed El Azzouzi, Gouenou Coatrieux, Reda Bellafqira, Denis Delamarre, Christine Riou, Naima Oubenali, Sandie Cabon, Marc Cuggia, and Guillaume Bouzillé. 2023. Automatic de-identification of french electronic health records: a cost-effective approach exploiting distant supervision and deep learning models. *BMC Medical Informatics and Decision Making*, 23(1):22.

European Parliament and Council of the European Union. General data protection regulation [online].

Daniele Faraglia and Other Contributors. 2010. Faker.

Andre Goncalves, Paroma Ray, Brendon Soper, et al. 2020. Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20(1):108.

Google Cloud. Deidentify sensitive data [online].

Imanol Isasa, Mikel Hernandez, Gorka Epelde, et al. 2024. Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis. *BMC Medical Informatics and Decision Making*, 24(1):27.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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.

Shreyasi Mandal and Ashutosh Modi. 2024. Iitk at semeval-2024 task 2: Exploring the capabilities of llms for safe biomedical natural language inference for clinical trials. *arXiv preprint arXiv:2404.04510*.

Stephanie M Meystre, Olivia Ferrandez, Jeff J Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Text de-identification for privacy protection: a study of its impact on clinical text information content. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Ministry of Electronics and Information Technology (MeitY), India. Indian digital personal data protection act 2023 [online].

Ministry of Health and Family Welfare (MoHFW), India. National resource centre for ehr standards (nrces) [online].

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Ishna Neamatullah, Margaret M Douglass, Li-wei H Lehman, Andrew T Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32.

PI-RoBERTa. Roberta-ner-personal-info [online].

PIB Press Release. Pib press release [online].

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Sunil Kumar Srivastava. 2016. Adoption of electronic health records: a roadmap for india. *Healthcare informatics research*, 22(4):261.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19.

Latanya Sweeney. 2013. Matching known patients to health records in washington state data. *arXiv preprint arXiv:1307.1370*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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 (2023). *arXiv preprint arXiv:2302.13971*.

U.S. Department of Health and Human Services (HHS). Health insurance portability and accountability act [online].

Matthijs J Warrens. 2015. Five ways to look at cohen's kappa. *Journal of Psychology and Psychotherapy*, 5.

Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(1):60.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Özlem Uzuner, Isaac Goldstein, Yuan Luo, and Isaac S. Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Appendix

Table of Contents

List of Tables

List of Figures

A Prompts and Synthetic Discharge Summaries

In Table 12, we showcase the prompts which we used to generate the ICDS_G^g and ICDS_G^l datasets. We used Prompt A in Table 12 for generating ICDS_G^g from Gemini-pro-1.0. Table 13 gives a sample discharge summary. Using prompt B in Table 12, we generated ICDS_G^l dataset using Llama-3-8B-Instruct. Table 14 gives a sample discharge summary.

B Tag Mapping across all the dataset and tag Distribution after Mapping the Tags

We have five datasets: n2c2-2006, n2c2-2014, ICDS_R, ICDS_G^g, and ICDS_G^l. Each dataset has its own tag set. n2c2-2006 contains 9 tags, n2c2-2014 contains 24, ICDS_R contains 26, ICDS_G^g contains 34, and ICDS_G^l contains 106 unique tags, including the `OTHERS` tag. In the datasets n2c2-2006, n2c2-2014, and ICDS_R, all the tags are related to PHI entities. However, in the ICDS_G^l and ICDS_G^g datasets, a few annotated tags are not related to the PHI entities due to LLM hallucinations. To train models for a fair comparison, we need a uniform tag set across all datasets.

Hence, we mapped the tag set of all the datasets to the n2c2-2006 tag set. In all the datasets, we mapped entities like street, city, country, zip, etc to `LOCATION`. Similarly, we mapped phone number, mobile number, email, landline, etc, to `CONTACT`. Additionally, we mapped all the PHI-related entities to their super-set using mapping shown in Table 15. In the ICDS_G^l and ICDS_G^g datasets, we have several tags unrelated to the PHI entities. Hence, we mapped all non-PHI entities to the `OTHERS` tag. After mapping the tag set of all the datasets to n2c2-2006 tag set, we calculated the tag distribution of all PHI entities across all datasets. The distribution of tag sets of all the dataset when mapped with n2c2-2006 dataset are shown in Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, and Fig. 12.

C Corpus Statistics

The n-gram frequencies from the n2c2-2006 dataset show a strong emphasis on clinical and procedural language, including terms like 'mg,' 'po,' and 'hospital,' as shown in Fig. 14. Notably, phrases such as 'discharge summary' and 'physical examination' dominate, highlighting standard documentation practices. Trigrams such as 'dis report status' and 'report status unsigned' indicate typical phrasing in medical reports. This is in contrast with the ICDS_R dataset in Fig. 18, where there is a predominance of time-related unigrams ('pm,' 'days') and clinical terms ('mgdl,' 'method'). The frequent bigrams and trigrams revolve around treatment and diagnosis descriptors, such as 'daily 10 days' and 'x ray chest,' illustrating the detailed recording of patient care routines and diagnostic procedures commonly found in medical records. In the n2c2-2006 dataset, bigrams like 'mg po' and 'discharge date,' and trigrams like 'mg po bid' and 'history present illness,' which reveal specific medication dosages and detailed descriptions of patient conditions, are found next to PHI elements, as shown in Fig. 16. In the ICDS_R dataset, specific trigrams like 'discharge summary crno' and 'normal discharge correspond' are located near PHI elements (Fig. 20). The differences between the n2c2 2006 dataset and ICDS_R highlight how clinical documentation practices and language differ between the US and India.

In the synthetic ICDS_G^g dataset, the frequent occurrence of 'phi' in various n-grams highlights (in Fig. 15) the inclusion of potentially identifiable information. Trigrams such as 'phi typehospitalfihphi' and 'phi typeid7673299w3phi' illustrate the use of placeholders for personal identifiers, indicative of the synthetic nature of the dataset and its focus on mimicking real-world PHI data while maintaining privacy. In the ICDS_G^l dataset, the frequent mention of basic terms like 'patient,' 'discharge,' and 'history' reflects their regular usage in clinical documents, as seen in Fig. 19. Phrases such as 'discharge summary' and 'medical history' indicate standardized document formats. For n-grams next to PHI elements in the synthetic ICDS_G^g dataset as seen in Fig. 17, we observe a mix of clinical terminology ('discharge,' 'patient,' 'history') and documentation descriptors ('text record,' 'reportend text'). Bigrams and trigrams like 'discharge summary patient' and 'text record record' suggest a replication of typical medical documentation formats. Terms like 'primary care physician' and 'history present illness' reflect the comprehensive nature of clinical narratives. In contrast, the n-grams next to PHI elements in the ICDS_G^l dataset, as shown in Fig. 21, highlight the frequent use of both temporal ('pm', 'days') and medical ('mgdl,' 'discharge') terms. Common bigrams and trigrams such as 'discharge summary,' 'cr x ray,' and 'x ray chest' underscore the clinical focus on diagnostic imaging and summary documentation. The

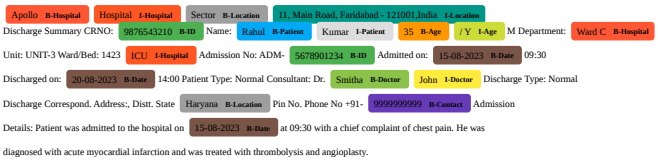


Figure 4: Pre-processed Discharge Summary after adding B and I tags

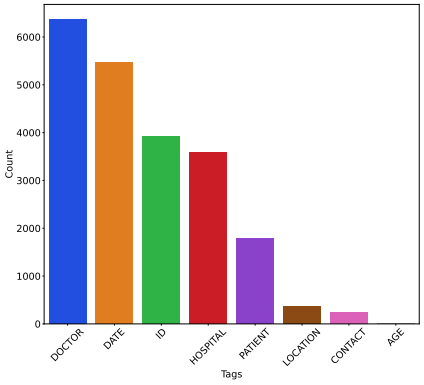


Figure 5: Tag Distribution in n2c2-2006 train dataset

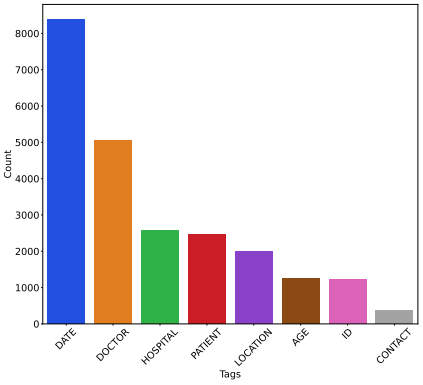


Figure 7: Tag Distribution in n2c2-2014 train dataset

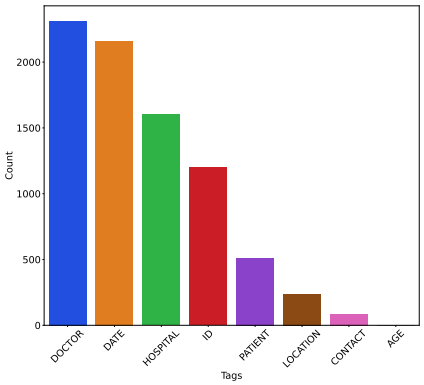


Figure 6: Tag Distribution in n2c2-2006 test dataset

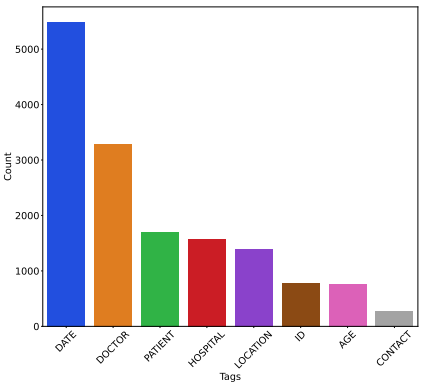


Figure 8: Tag Distribution in n2c2-2014 test dataset

trigrams involving 'daily 10 days' and 'x ray chest bed' reflect specific medical interventions and patient care protocols typically documented in patient records.

D Model Training Details

We fine-tuned dslim/bert-base-NER (Dslim bert base NER), ghadeermobasher/BCHEM4-Modified-BioBERT-v1 (BioBERT), and Clinical-AI-Apollo/Medical-NER (Clinical AI Apollo). We obtained a consistent train-set F1 Score for PHI entities from these models after fine-tuning, but the performance of these models decreased significantly when we tested them on cross-dataset settings. However, after fine-tuning, PI-RoBERTa outperformed these models in the same and cross-dataset settings, so we chose PI-RoBERTa for further experiments. Fig. 13 shows the model architecture.

PI-RoBERTa was fine-tuned on each training set as given in Table 3 and tested on each corresponding test set. We fixed the hyperparameters for all the experiments. The model was fine-tuned at four epochs in all the experiments with a batch size of 8; the learning rate was 5e-5. We used Weighted Cross entropy loss to handle the data imbalance problem because around 90 percent of the tokens correspond to non-PHI entities in all datasets. After several experiments, we devised a formula to assign weights to different Entities. $w_t = \log\left(4 \times \frac{n}{n_t}\right)$, where w_t is the weight assigned to the t^{th}entity; n_t is the number of tokens in the t^{th} entity; n is the total number of tokens in the dataset

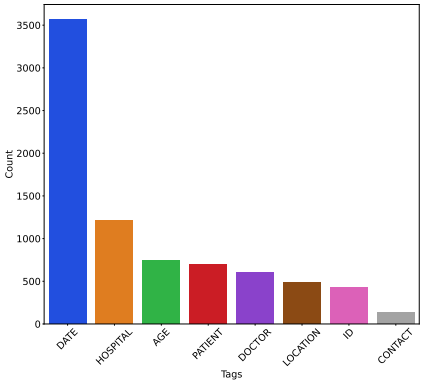


Figure 9: Tag distribution in ICDS$_R$ train dataset

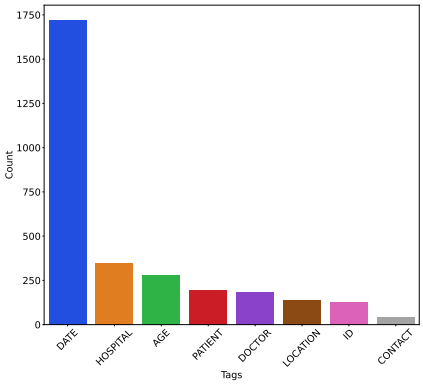


Figure 10: Tag distribution in ICDS$_R$ test dataset

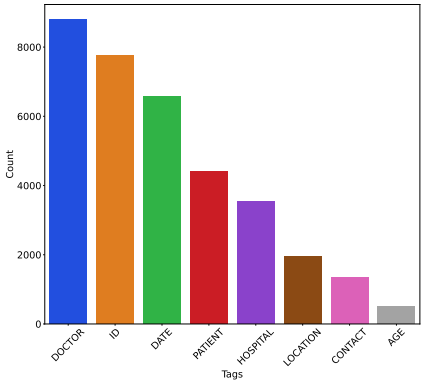


Figure 11: Tag distribution in ICDS^g_G dataset

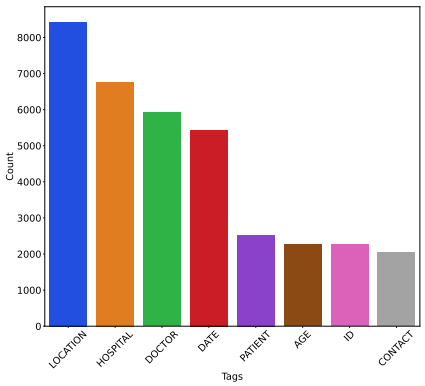


Figure 12: Tag distribution in ICDS^l_G dataset

E Evaluation Metrics

Model was evaluated using various performance metrics as described below.

- Macro Precision:

$$\text{Precision}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i}$$

- Macro Recall

$$\text{Recall}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i}$$

- Macro F1-score

$$\text{F1-score}_{\text{macro}} = \frac{2 \times \text{Precision}_{\text{macro}} \times \text{Recall}_{\text{macro}}}{\text{Precision}_{\text{macro}} + \text{Recall}_{\text{macro}}}$$

- Micro Precision:

$$\text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)}$$

- Micro Recall

$$\text{Recall}_{\text{micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)}$$

- Micro F1-score

$$\text{F1-score}_{\text{micro}} = \frac{2 \times \text{Precision}_{\text{micro}} \times \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}$$

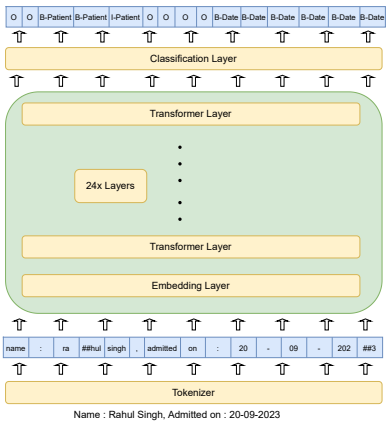


Figure 13: Architecture of PI-RoBERTa

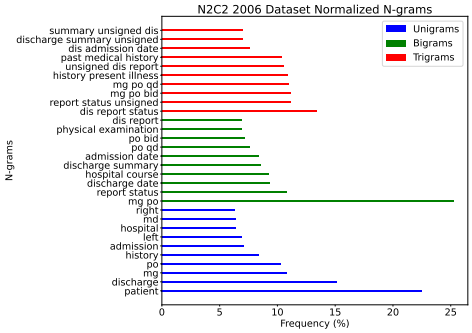


Figure 14: n2c2-2006 Top 10 N-grams

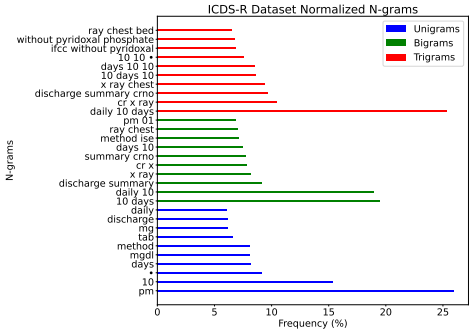


Figure 18: ICDS$_R$ Top 10 N-grams

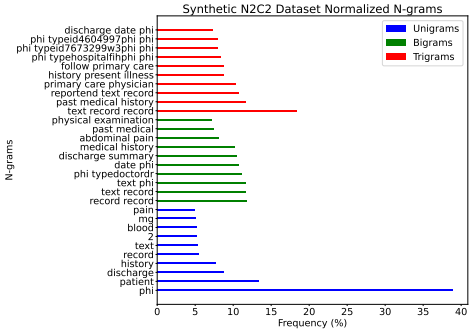


Figure 15: ICDS^g_G Top 10 N-grams

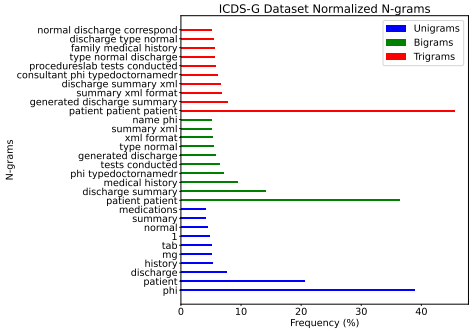


Figure 19: ICDS^l_G Top 10 N-grams

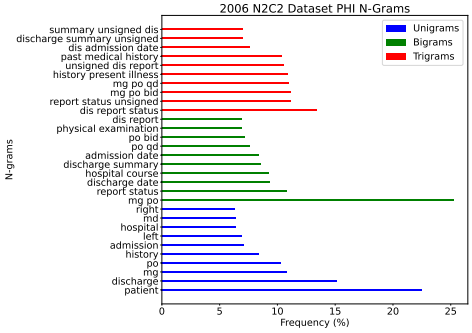


Figure 16: n2c2-2006 Top 10 N-grams Spanning PHI Elements

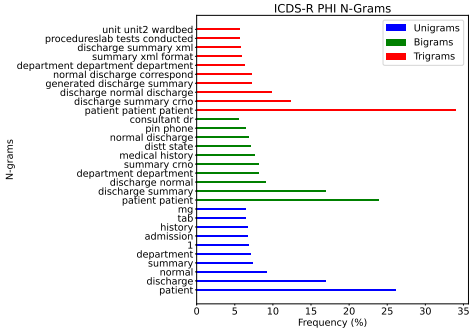


Figure 20: ICDS$_R$ Top 10 N-grams Spanning PHI Elements

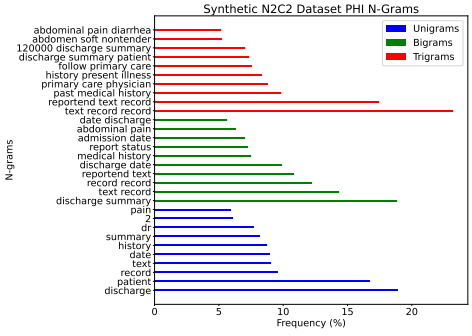


Figure 17: ICDS^g_G Top 10 N-grams Spanning PHI Elements

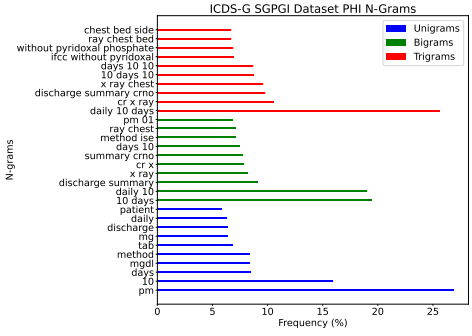


Figure 21: ICDS^l_G Top 10 N-grams Spanning PHI Elements

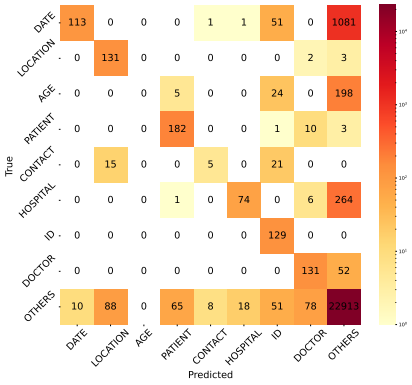


Figure 22: Confusion matrix on ICDS$_R$ test set when PI-RoBERTa finetuned on n2c2-2006

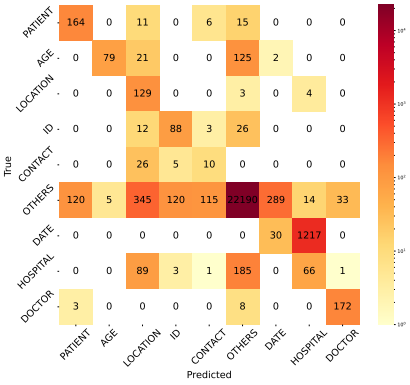


Figure 25: Confusion matrix on ICDS$_R$ test set when PI-RoBERTa finetuned on ICDS^g_G

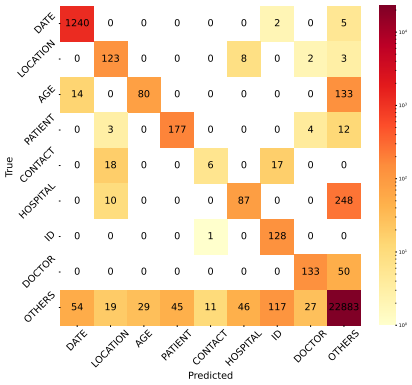


Figure 23: Confusion matrix on ICDS$_R$ test set when PI-RoBERTa finetuned on n2c2-2014

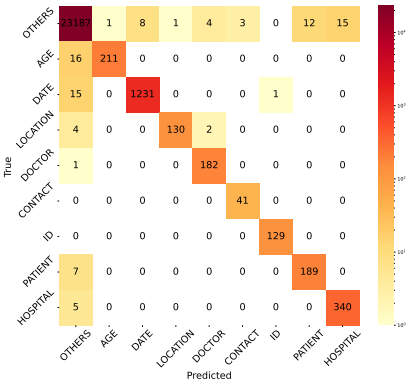


Figure 26: Confusion matrix on ICDS$_R$ test set when PI-RoBERTa finetuned on ICDS^l_G

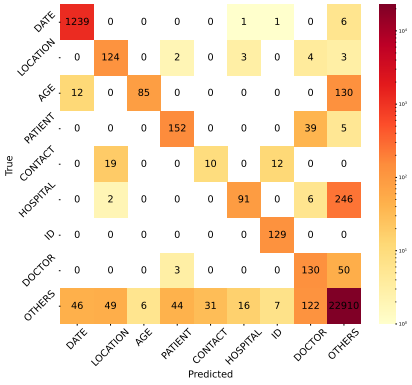


Figure 24: Confusion matrix on ICDS$_R$ test set when PI-RoBERTa finetuned on Combining n2c2-2006 and n2c2-2014

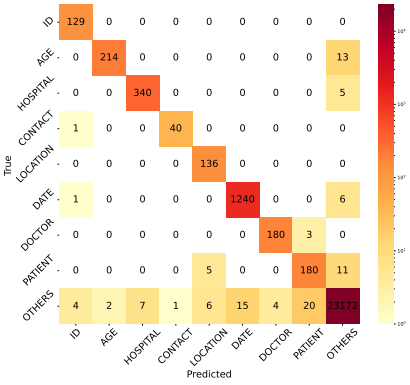
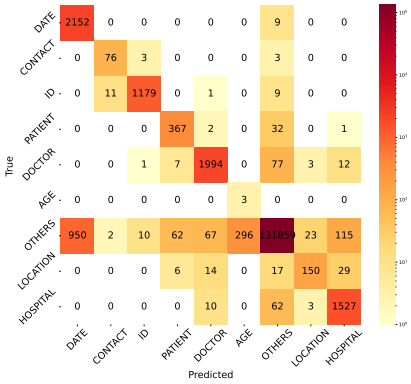
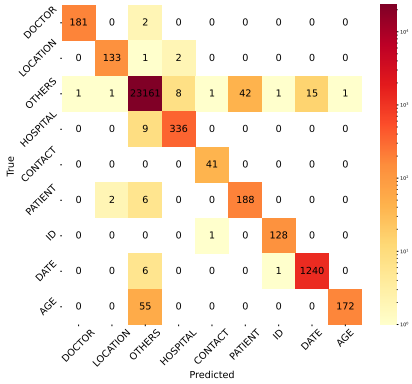


Figure 27: Confusion matrix on ICDS$_R$ test set when PI-RoBERTa finetuned on Combining ICDS^l_G and ICDS^g_G dataset



(a) Confusion matrix on n2c2-2006 test set when PI-RoBERTa finetuned on combining n2c2-2014, ICDS^g_G, and ICDS^l_G dataset.



(b) Confusion matrix on ICDS$_R$ test set when PI-RoBERTa finetuned on combining ICDS^l_G, ICDS^g_G, n2c2-2006, and n2c2-2014 dataset.

Figure 28: Confusion matrix on n2c2-2006 and ICDS$_R$ testset when PI-RoBERTa fine-tuned on combination of generated and real data.

Prompt Id	Prompt
A	Generate discharge summaries for Indian patients, capturing the essence of healthcare in India. The summaries should integrate conventional medical treatments with traditional remedies, reflecting the holistic approach embraced by Indian healthcare systems. Incorporate prevalent Indian health conditions, treatments, and culturally relevant follow-up care instructions. To ensure authenticity, each summary should include distinct patient details like name, age, address, contact information, hospital, doctor, and ID. Include prevalent diseases in India such as Tuberculosis (TB), Diabetes, Cardiovascular Diseases, Respiratory Infections, Hypertension, Dengue Fever, Malaria, Hepatitis, Chronic Kidney Disease (CKD), Cancer, Typhoid Fever, Cholera, HIV/AIDS, Japanese Encephalitis, Leptospirosis, Rabies, Tuberculosis of the Central Nervous System (CNS TB), Rheumatic Heart Disease, Iron-Deficiency Anemia, and Chikungunya. Also, laboratory test reports of the chosen disease should be included. Ensure the format of generated discharge summaries is similar to the summary given in the prompt, i.e., in XML format. Example: Patient Summary: <discharge summary> Generate the summaries that have a minimum of 2048 words. Ensure there is consistent consistency between the doctor's name, patient name, drug-disease, etc.
B	Generate an extensive discharge summary of at least 2048 words tailored for Indian patients. To ensure authenticity, the generated summary must include distinct patient-specific details like name, age, address, contact information, hospital name, doctor name, and unique ID. Maintain coherence across all the elements, doctor's name, patient's identity, medications, diseases, etc. Ensure all the PHI (personal health information) elements are properly annotated to maintain privacy and authenticity. The generated discharge summary should be XML-formatted with PHI annotations. The generated summaries should include following sections: Admission Details, Diagnosis / Chief Complaints, Allergies, Physical Examination, Medical History, Family Medical history, Treatment Plan, Investigations, Medications (List of medications prescribed at discharge), Follow-up Instructions, Procedures/Lab Tests Conducted (List of procedures or tests conducted during hospital stay, along with results if available), and Special Instructions. Please ensure that these sections are incorporated into the generated summaries, but refrain from including them as tags in the output. The generated summary should be properly enclosed within the <RECORD> and </RECORD> tags to ensure it's within the XML format. Here's an example patient summary: Patient Summary: <discharge summary>
C	Generate an extensive synthetic discharge summary of at least 2048 words tailored for Indian patients. Generated summary must include distinct entities like name, age, address, contact information, hospital name, doctor name, and unique ID. Maintain coherence across all the elements, doctor's name, patient's identity, medications, diseases, etc. Identify all entities in the generated text and mark these with XML tags as in the following example:<TYPE='Insurance Number'>AB123456C</TYPE> entities= ['Patient Name','Hospital_Name','Staff_Name','Doctor_Name','Age','Gaurdian_Name','Gender', 'Patient_ID','Misc_Medical_ID','Aadhar','Driver_License','Voter_ID','PAN_Card','Patient_DOB', 'Treatment_Date','Treatment_Time','Phone_No','Landline','Email','IP_Address','Fax','Doctor_Specialisation', 'Patient_Profession','City','Ward_Location','Device_Number','Other_Info','State','Street','Zip','Country', 'Other_Location','Other_Govt_ID','Insurance_Number','Web_url'] The generated summaries should include following sections: Admission Details, Diagnosis / Chief Complaints, Allergies, Physical Examination, Medical History, Family Medical history, Treatment Plan, Investigations, Medications (List of medications prescribed at discharge), Follow-up Instructions, Procedures/Lab Tests Conducted (List of procedures or tests conducted during hospital stay, along with results if available), and Special Instructions. Please ensure that these sections are incorporated into the generated summaries, but refrain from including them as tags in the output. The generated summary should be properly enclosed within the <RECORD> and </RECORD> tags to ensure it's within XML format. Here's an example synthetic patient summary: <discharge summary>.

Table 12: Prompts used for synthetic discharge summary generation

Model Name	Generated Discharge Summary
Gemini-pro-1.0	829369 HODYEGE DISTRICT HOSPITAL 9151843 830690 408542 12/12/10/2023 23:59:00 PM Discharge Summary Discharge Report Status : Signed DISCHARGE SUMMARY NAME : NIVEDITHA CHETTY UNIT NUMBER : 393-29-55 ADMISSION DATE : 12/10/10/2023 DISCHARGE DATE : 12/12/10/2023 PRINCIPAL DIAGNOSIS : Prolapsed intervertebral disc with radiculopathy of L4-L5, L5-S1, L2-L3, with weakness of right lower limb. ASSOCIATED DIAGNOSIS : CVA (LEFT SIDED HEMIPARESIS) ADVERSE DRUG REACTIONS : None . CLINICAL INFECTIONS : None . PRINCIPAL PROCEDURE : Spinal surgery ASSOCIATED PROCEDURES : Physiotherapy . HISTORY OF PRESENT ILLNESS : Mrs. NIVEDITHA CHETTY is a 58 years old female who presented to the emergency department with complaints of low back pain for 1 year and weakness of right lower limb for 1 month . Weakness was insidious in onset and gradually increased to the present level . There was no numbness of lower limbs . No urinary or bowel involvement . No fever , cough , gastrointestinal symptoms or cardio-respiratory symptoms . PAST MEDICAL HISTORY : CVA (LEFT SIDED HEMIPARESIS) . SOCIAL HISTORY : Non smoker and non alcoholic . PHYSICAL EXAMINATION : VITALS : Temperature : 98.6\302\260 F, BP : 120/80 mmHg , SPO2 : 95% , Respiratory rate : 20/min , Pulse rate : 80/min CNS : GCS 15/15 , Cranial nerves - Normal . Power - Right lower extremity grade - 4/5 . Sensation - Intact . Motor system examination - Tone - Increased on the right side . Deep tendon reflexes - Brisk on the right side . Plantar - Extensor on the right side . CVS : No pallor , icterus , clubbing , pedal edema , cyanosis , SPO2 - 99% . JVP not raised . Carotid pulsations are well felt and equal on both the sides . No bruits audible . Heart sounds - Regular . No murmurs , gallops , rubs or extra sounds . Respiratory system : Chest expansion - Symmetrical . Air entry reduced bilaterally but more on the right side . No added sounds . Breath sounds decreased in right infrascapular area . Abdomen : Soft , non-tender , no mass or organomegaly . LABORATORY DATA : Hb - 12.0 gm/dl , TLC - 8,000/cmm , Platelets - 2.5 lacs/cmm , ESR - 20 mm/hr . Blood sugar - 100 mg/dl , Urea - 30 mg/dl , Creatinine - 1.0 mg/dl , Uric acid - 5.0 mg/dl . Liver function tests - Normal . Lipid profile - Normal . Electrolytes - Normal . X-RAY : LUMBAR SPINE AP LATERAL VIEW : There is narrowing of the L4-L5, L5-S1, L2-L3 intervertebral spaces with sclerosis of the adjacent end plates . Loss of lumbar lordosis with mild dextroscoliosis . CT SCAN : LUMBAR SPINE : There is a prolapsed intervertebral disc at L4-L5, L5-S1, L2-L3 with foraminal stenosis on the right side at L4-L5 . Thecal sac is compressed at L4-L5, L5-S1 due to central disc prolapse . MRI SCAN : LUMBAR SPINE : There is a prolapsed intervertebral disc at L4-L5, L5-S1, L2-L3 with foraminal stenosis on the right side at L4-L5 . Thecal sac is compressed at L4-L5, L5-S1 due to central disc prolapse . There is evidence of denervation of the right sciatic nerve and cauda equina . HOSPITAL COURSE : The patient was admitted for surgical intervention . After proper pre-operative evaluation the patient was taken up for surgery . During surgery , a decompression of the thecal sac and foramina was done . Postoperatively , the patient was started on physiotherapy . The patient was able to walk with the support of walker . MEDICATIONS ON DISCHARGE : Tramadol - 50 mg - BD SOS - 7 days Aceclofenac - 100 mg - BD - 7 days Multivitamin and minerals - OD FOLLOW UP : The patient will be reviewed in the clinic in 2 weeks . DR.GYANESHWARI JANGID MBBS, MD, MS (Orthopedics) HODYEGE DISTRICT HOSPITAL Electronically Signed DR.GYANESHWARI JANGID [report_end]

Table 13: Example summary generated using gemini-pro-1.0

Model Name	Generated Discharge Summary
Meta-Llama-3-8B-Instruct	Max Super Specialty Hospital Sector 11, Dwarka, New Delhi - 110075,India Discharge Summary CRNO: 1234567890 Name: Rahul Kumar 35/Y/M Department: Ward E Unit: UNIT-3 Ward/Bed: 5432 ICU Admission No: ADM-9012345678 Admitted on: 25-08-2023 14:30 Discharged on: 28-08-2023 10:00 Patient Type: Normal Consultant: Dr. Rohan Sharma Discharge Type: Normal Discharge Correspond. Address:, Distt. State Uttar Pradesh Pin No. Phone No +91-9812345678 FOLLOW UP: Appointment on 15/10/2023 Morning Block-Upto 09:00 A.M.INVESTIGATION RESULTS: 25/08/2023 09:30 AM Blood Urea 40 mg/dL 25/08/2023 09:30 AM Serum Creatinine 1.2 mg/dL 25/08/2023 09:30 AM Hb 12.5 g/dL Admission Details: Patient was admitted to the hospital with a diagnosis of Acute Kidney Injury (AKI) due to dehydration. He was a known hypertensive and diabetic patient. He presented with symptoms of fatigue, weakness, and decreased urine output. Diagnosis/Chief Complaints: Primary Diagnosis: Acute Kidney Injury (AKI) Secondary Diagnosis: Hypertension, Diabetes Mellitus Allergies: None reported Physical Examination: The patient was conscious and oriented. He had a pulse rate of 100 beats per minute, blood pressure of 160/100 mmHg, and respiratory rate of 20 breaths per minute. He had bilateral pedal edema and decreased urine output. Medical History: The patient had a history of hypertension and diabetes mellitus. He was taking medications for these conditions, including metformin and lisinopril. Family Medical History: The patient's father had a history of hypertension and heart disease. Treatment Plan: The patient was treated with intravenous fluids and medications to manage his symptoms. He was also started on dialysis to manage his acute kidney injury. Investigations: The patient underwent the following investigations: Blood Urea 40 mg/dL Serum Creatinine 1.2 mg/dL Hb 12.5 g/dL Medications: The patient was prescribed the following medications at discharge: Metformin 500 mg twice daily Lisinopril 10 mg daily Furosemide 20 mg daily Hydrochlorothiazide 25 mg daily Aspirin 75 mg daily Follow-up Instructions: The patient was advised to follow up with his primary care physician in two weeks for a check-up and to monitor his blood sugar and blood pressure levels. Procedures/Lab Tests Conducted: The patient underwent the following procedures and lab tests during his hospital stay: Blood tests: Blood Urea 40 mg/dL, Serum Creatinine 1.2 mg/dL, Hb 12.5 g/dL Urine tests: Urine output decreased, proteinuria Dialysis: Three sessions of dialysis were performed Special Instructions: The patient was advised to follow a strict fluid intake and to monitor his urine output closely. He was also advised to take his medications as prescribed and to follow up with his primary care physician regularly. Prepared by (Dr. Rakesh Kumar) Signature of Consultant Printed on 29-08-2023 14:30:15 Temp Employee Male @ 192.168.1.100 Page 2 / 2

Table 14: Example summary generated using llama-3-8B-Instruct

Original tags	Mapped Tags
Treatment_Date, Patient_DOB, Investigation_Date, Admission Date, Procedure_Date, Date	DATE
Ward_Location, Hospital_Name, Department	HOSPITAL
Patient_ID, Misc_Medical_ID, Employee_ID, Admission Number	ID
Age	AGE
Doctor_Name, Staff_Name, Prepared by, Signature, Doctor_Signature, Signature of Consultant	DOCTOR
Patient_Name, Gaurdian_Name, Patient_Signature, Patient_Spouse, Family_Member_Name	PATIENT
Zip, Phone_No, Landline, IP_Address, Phone, Contact_Info, Contact_Number, Contact_No, Mobile, Phone Number, Patient_Phone, Email, Email_ID, Contact Information, Phone No	CONTACT
City, State, Country, Street, Other_Location, Correspondence_Address, Contact_Address, Contact Information, Pin, Pin Code, Pin_No, Postal_Code, Address, Contact_Address	LOCATION

Table 15: Tag mapping from PHI entities in the different datasets to the PHI entity set of n2c2-2006 dataset, and all other non-PHI entities are mapped with `Others` tag