Towards Fair and Efficient De-identification: Quantifying the Efficiency and Generalizability of De-identification Approaches

Noopur Zambare[1] **Kiana Aghakasiri**[1] **Carissa Lin**[1]
Carrie Ye[2,4] **J. Ross Mitchell**[1,2,3] **Mohamed Abdalla**[1,2,3]

[1]Department of Computing Science, University of Alberta
[2]Department of Medicine, University of Alberta
[3]Alberta Machine Intelligence Institute (Amii)
[4]Arthritis Research Canada

zambare, kaghakas, carissa1, cye, jmitche2, mabdall2@ualberta.ca

Abstract

Large language models (LLMs) have shown strong performance on clinical de-identification, the task of identifying sensitive identifiers to protect privacy. However, previous work has not examined their generalizability between formats, cultures, and genders. In this work, we systematically evaluate fine-tuned transformer models (BERT, ClinicalBERT, ModernBERT), small LLMs (Llama 1-8B, Qwen 1.5-7B), and large LLMs (Llama-70B, Qwen-72B) at de-identification. We show that smaller models achieve comparable performance while substantially reducing inference cost, making them more practical for deployment. Moreover, we demonstrate that smaller models can be fine-tuned with limited data to outperform larger models in de-identifying identifiers drawn from Mandarin, Hindi, Spanish, French, Bengali, and regional variations of English, in addition to gendered names. To improve robustness in multi-cultural contexts, we introduce and publicly release BERT-MultiCulture-DEID, a set of de-identification models based on BERT, ClinicalBERT, and ModernBERT, fine-tuned on MIMIC with identifiers from multiple language variants. Our findings provide the first comprehensive quantification of the efficiency-generalizability trade-off in de-identification and establish practical pathways for fair and efficient clinical de-identification.

Details on accessing the models are available at: https://doi.org/10.5281/zenodo.18342291

1 Introduction

De-identification is the process of removing personally identifiable information (PII) from data to protect individual privacy. This step is crucial in healthcare research, where clinical texts often contain sensitive details such as patient names, addresses, contact information, and other identifiers. Regulations like the Health Insurance Portability and Accountability Act (HIPAA) (Centers for Medicare & Medicaid Services, 1996) and the General Data Protection Regulation (GDPR) (European Union, 2016) require the de-identification of such clinical notes before sharing. However, protecting this sensitive information is challenging, leading healthcare organizations to restrict sharing such data.

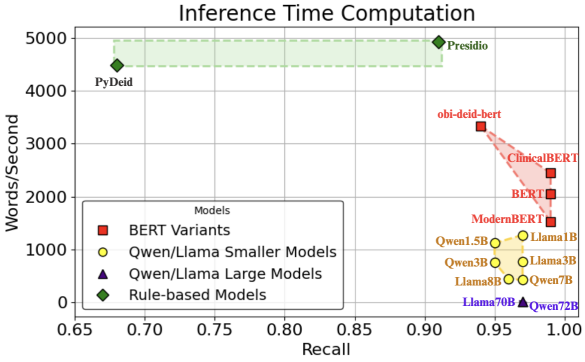

Figure 1: Inference time computation of different de-identification models. The y-axis shows words processed per second, and the x-axis shows recall. Results are grouped by model type: BERT variants, smaller Qwen(1.5-7B)/Llama(1-8B) models, large Qwen-72B/Llama-70B models, and rule-based models.

Traditionally, de-identification was performed using dictionary or rule-based systems, which rely on predefined patterns or lists of dictionaries to identify PII. Although these methods are computationally inexpensive, they often fail to generalize between note formats or variations in identifier style, particularly when dealing with identifiers from different cultural contexts or institutions (Fer-

Findings of the Association for Computational Linguistics: EACL 2026, pages 4242–4257

rández et al., 2012).

To improve the generalizability of de-identification, researchers have adopted transformer-based models such as BERT (Devlin et al., 2019) and ClinicalBERT (Alsentzer et al., 2019). These models have demonstrated substantial improvements in de-identification performance. Using contextual embeddings, these models can detect sensitive tokens with high precision and recall, even in complex narrative structures. However, models like BERT require fine-tuning for each type of identifier (e.g., names, addresses, phone numbers), and this process requires labeled data.

To minimize the need for training data (generated by the data owners looking to de-identify their data), researchers have more recently experimented with using LLMs such as Llama (Grattafiori et al., 2024), Qwen (Bai et al., 2023; Team, 2024), Mixtral (Jiang et al., 2024), and ChatGPT (Radford et al., 2018). Recent work suggests that these models perform well if not better than BERT-based models (Altalla' et al., 2025; Wiest et al., 2025; Pissarra et al., 2024).

Although LLM-based de-identification has demonstrated high performance, certain critical challenges limit its adoption by healthcare institutions. First, most academic research on LLM-based de-identification experiment with very large models, which are difficult to deploy locally for many institutions. Furthermore, even if institutions had the computational ability to run these models, inference with large models remains computationally intensive and slow, making them impractical for clinical deployment (though this has not been formally quantified in any previous work). Given the constraints of real-world clinical settings, understanding this trade-off is crucial.

At the same time, while some studies have examined cross-format or cross-lingual performance, comprehensive evaluations of LLM-based de-identification models across culturally diverse identifiers, different formats, and gendered identifiers are lacking, limiting understanding of their robustness and real-world applicability.

Taken together, these gaps raise two critical questions: 1) How (in-)efficient is LLM-based de-identification and can smaller LLM de-identification models achieve performance comparable to very LLMs (e.g. 70B parameters) while substantially reducing computational requirements and inference time?, and 2) Are LLM-based de-identification models able to generalize across healthcare institutions, genders and language/cultural variations in identifiers?

To address these questions, our work makes the following contributions:

- **Efficient De-identification:** We conduct a systematic analysis of inference efficiency and performance across fine-tuned BERT models (BERT, ClinicalBERT, ModernBERT) as well as smaller Llama (1B, 3B, 8B) and Qwen (1.5B, 3B, 7B) models. We demonstrate performance comparable to large models (Llama 3.3-70B and Qwen2.5-72B) with substantially lower compute requirements and faster inference time.

- **Generalizability:** We evaluate model robustness across formats, gendered identifiers, and culturally diverse identifiers drawn from Mandarin, Hindi, Spanish, French, and English naming traditions. Our goal is to test whether English models remain reliable when the identifiers reflect the cultural diversity seen in large urban anglophone cities. Our results highlight the limitations of all models in adapting to diverse real-world contexts.

- **Multi-Cultural Deid:** To address substantial performance degradation in multicultural identifiers, we fine-tuned BERT, ClinicalBERT, and ModernBERT, developing BERT-MultiCulture-DEID, a set of models that exhibit improved multi-cultural generalization.

2 Related Work

De-identification has been an active field of research for over 30 years (Sweeney, 1998). During this time, the field has continuously adopted the most advanced methods, from regular expressions (Neamatullah et al., 2008) to traditional machine learning algorithms (Liu et al., 2017), and transformer-based models (e.g., BERT and ClinicalBERT) (Johnson et al., 2020). Consequently, the performance of automated de-identification systems has continued to improve with performance metrics (e.g., precision and recall) nearing perfection (Johnson et al., 2020; Moore et al., 2023). Most recently, LLMs (e.g., Llama, Qwen, Mixtral) have been adopted for de-identification (Pissarra et al., 2024; Altalla' et al., 2025; Wiest et al., 2025) and have demonstrated near-perfect recall without

requiring (much) labeled training data, demonstrating impressive zero-shot and few-shot generalization (Brown et al., 2020).

2.1 Efficiency in Clinical De-identification

De-identification is a time-consuming task for humans. Dorr et al. (2006) measured that de-identifying a clinical note required 87.3 ± 61 seconds, while Douglass et al. (2004) found that humans were able to de-identify between 250 and 350 words per minute.

Heider et al. (2020) conducted a comparative evaluation of de-identification systems (Amazon Comprehend Medical PHId (Amazon Web Services, 2018), Clinacuity (Meystre et al., 2023), and the National Library of Medicine's (NLM) Scrubber (National Library of Medicine, 2019) and concluded that none of the systems simultaneously achieved optimal performance in both speed and accuracy. NLM Scrubber achieved a recall of 0.47 and processed 4.09 notes/sec, CliniDeID achieved a recall of 0.99 at 0.29 notes/sec, and Amazon Comprehend Medical PHId achieved a recall of 0.80 at 1.75 notes/sec.

As de-identification models have grown more intricate, their inference time has also increased. To address increases in compute time, Sundrelingam et al. (2025) introduced pyDeid for rapid, generalizable rule-based de-identification, showing competitive speed with an average runtime of 0.48 seconds/note with a best recall of 0.95, outperforming traditional tools such as Deid (0.93 seconds/note, recall: 0.87; (Neamatullah et al., 2008)) and Philter (6.38 seconds/note, recall: 0.92; (Norgeot et al., 2020)). Other researchers have sought to enhance the efficiency of LLMs by fine-tuning smaller models. Dorémus et al. (2025) demonstrated that small LLMs (e.g., 7B) can be fine-tuned to achieve high de-identification performance (F1 score: 0.97 and recall: 0.93), although they did not specifically measure inference time. Chen et al. (2025) and Naguib et al. (2024) evaluated pretrained models on biomedical NLP tasks (performance and cost) and NER (performance and carbon emissions), respectively.

Current literature exhibits several shortcomings. First, existing efficiency assessments use different datasets, which hinders direct comparability. Second, these evaluations lack uniformity in metrics; measuring time per note complicates future comparisons due to variability in note length. Finally, there have been no direct comparisons of the efficiency of LLMs with other competing models.

2.2 Generalization and Robustness

Recent advances in pre-trained transformer-based models, such as variants of BERT and GPT, Llama, etc., (Radford et al., 2018) have demonstrated strong generalization capabilities in various NLP tasks (Budnikov et al., 2025). BERT-based models typically excel in domain-specific fine-tuning due to their bidirectional contextual representations, whereas GPT-style models benefit from autoregressive pre-training that allows effective few-shot and zero-shot generalization (Brown et al., 2020).

Recognizing potential concerns about robustness, researchers have explored the generalizability of clinical de-identification models. Xiao et al. (2023) identified significant performance disparities in most de-identification approaches they evaluated, with biases evident in demographic dimensions such as gender and race. Chen et al. (2024) analyzed pre-trained transformer models in discharge summaries and nursing notes, discovering significant accuracy declines due to differences in structure and PII entity distributions. Kim et al. (2024) observed that state-of-the-art de-identification models show poor generalization on new datasets, mainly due to challenges in maintaining training corpora, as well as variations in labeling standards and patient record formats across institutions. They also suggested that GPT models (and large LLMs more generally) could help address these challenges, an assumption which our analysis shows is incorrect.

Unfortunately, current research has several limitations. First, assessments are frequently confined to smaller models, as illustrated by Chen et al. (2024), who restricted their evaluation to BERT-based models. Second, previous studies often limit their evaluations to particular types of PII. For example, Xiao et al. (2023) focused solely on the robustness of the models with respect to patient names. However, in practice, other identifiers, such as addresses (including postal codes and street patterns), institution or hospital names, phone numbers, and other entities, also vary across different contexts.

Finally, while these studies perform fine-tuning of transformer-based models for de-identification, they do not analyze how fine-tuning impacts robustness in smaller models across formats, genders, and identifiers drawn from different language variants.

3 Datasets

We performed experiments using two English-language clinical datasets: the publicly available MIMIC-III dataset (Johnson et al., 2016) and a smaller proprietary dataset consisting of rheumatology referral letters. In this work, we define PII as any token belonging to one of the following categories: name, phone/fax, hospital, city, state, address, country, company, university, date, email, and other (MRN, account number, etc.). Appendix Table 4 presents some descriptive statistics of the two datasets, and Figure 5 shows the statistics of identifiers in the test data.

All datasets, codes and models utilized in this study were collected and used in compliance with their respective licenses and access requirements.

3.1 MIMIC-III

We targeted PII identified in the Health Insurance Portability and Accountability Act (HIPAA) for MIMIC-III. These include patient names, telephone numbers, addresses, and dates. We sampled 4,000 discharge summaries from the MIMIC-III dataset. Of these, 2,000 were used for fine-tuning with varying subset sizes (250, 500, 1,000, and 2,000 notes), 1,000 were used for validation, and 1,000 were sequestered for testing.

3.2 Private Clinical Dataset (PCD) - Alberta Health Services

This dataset consists of 204 referral letters from physicians to rheumatologists, covering a wide range of clinical scenarios. All notes were saved in PDF format and consisted of a combination of digitally typed documents, scanned files, and handwritten letters. We extracted text from these PDFs using optical character recognition (OCR) with the Doctr library (Mindee, 2021). Authors then manually removed all sensitive information, including patient names, provider names, phone numbers, addresses, medical record numbers, account numbers, other identifiers, dates, and any additional personally identifiable information. The collection and use of this dataset was approved by the University of Alberta's REB (#Pro00141020).

3.3 Faker

We substituted the de-identified masks in the original notes with realistic surrogate data, preserving the structure and readability of the clinical text while ensuring that no real patient information was exposed. Replacement was performed using the Python-based Faker library (Faraglia and Other Contributors, 2014). This library provides a wide variety of synthetic identifiers, including names, addresses, phone numbers, email addresses, organization names, dates, and other identifiers such as medical record numbers and account numbers. For all experiments, except where specified, we used the default settings (e.g., using the US-locale).

4 Models

We tested multiple models: pretrained models including Llama 3.1-8b, Llama 3.2-1B, Llama 3.2-3B, Llama 3.3-70B, and Qwen 2.5 (1.5B, 3B, 7B, and 72B), BERT, ClinicalBERT, ModernBERT, and three open source toolkits (obi-deid-bert (OBI Organization, 2022), pyDeid (Sundrelingam et al., 2025), and Presidio (Microsoft, 2018)).

4.1 BERT, ClinicalBERT and ModernBERT

BERT is a transformer-based language model pretrained on large corpora using a masked language modeling objective, enabling it to capture rich contextual representations of text (Devlin et al., 2019). ClinicalBERT is a variant of BERT that is further pre-trained on clinical notes and discharge summaries from the MIMIC III dataset (Huang et al., 2019). This improves ClinicalBERT's ability to identify medical terminology and context. ModernBERT (Warner et al., 2024) is a more recent adaptation of BERT, offering better performance, efficiency, and longer sequence handling compared to standard BERT models.

In the experiments below, we fine-tuned various BERT, ClinicalBERT, and ModernBERT models as comparators. In this section, we describe generalized fine-tuning information used for all varieties. The specifics for each variation are described in their respective sections. The models were fine-tuned to perform binary token classification or multiclass token classification. For binary classification, each token was classified as PII or non-PII. For multiclass classification, each token was classified as non-PII or as a member of a set with multiple PII types. We focus on the results of binary PII prediction, with multiclass results presented in Appendix D.

To ensure a fair comparison across different dataset sizes (250, 500, 1000, and 2000 samples), the number of training epochs was kept the same for all models. We selected 10 epochs because of

stabilization in validation loss at this number.

We also included *obi-deid-bert*(OBI Organization, 2022), a publicly available de-identification model based on ClinicalBERT (pre-trained on MIMIC-III and fine-tuned on i2b2/n2c2 (Stubbs and Uzuner, 2015)), as a baseline.

4.2 LLMs (Llama and Qwen)

LLMs are transformer-based models pretrained using the next-token prediction objective, giving rise to generative capabilities. Llama (Grattafiori et al., 2024) and Qwen (Bai et al., 2023) are families of open-weight language models.

We fine-tuned the Llama (1B, 3B, 8B) and Qwen (1.5B, 3B, 7B) models for token classification. In this work, we used them exclusively for token-level classification. A classification head was appended to the final transformer layer, and fine-tuning was restricted to Low-Rank Adaptation (LoRA) and the classification head. Small models (1B/1.5B) were fine-tuned for 5 epochs, medium-sized models (3B) for 5 epochs, and large models (7B and 8B) for 3 epochs. Detailed experimental setup is explained in the Appendix C.

4.3 Prompt-tuning

We prompt-tuned the largest variants, Llama-70B and Qwen-72B, using one-shot prompting on a small subset of five MIMIC-III notes, using the prompt mentioned in Appendix Figure 7, detalied in Appendix Section E.

4.4 Presidio

Microsoft Presidio (Microsoft, 2018), an open source framework, detects PII using named entity recognition (NER) techniques.

4.5 PyDeid

We evaluated PyDeid (Sundrelingam et al., 2025), a rule-based system that uses regular expressions and fixed inclusion/exclusion lists to de-identify free-text clinical data.

5 Metrics

5.1 Standard Classification Metrics

For the majority of our analyses, we present precision and recall – standard classification metrics used in de-identification. Precision is the proportion of tokens predicted as PII that are actually PII. Recall is the proportion of PII in a note that is correctly flagged as PII.

Model	P	R	CIRE	Time (s)	STD (s)	Words/sec
BERT	0.99	0.99	0.99	67	2.0	2048
ModernBERT	0.99	0.99	0.99	90	3.1	1528
ClinicalBERT	0.99	0.99	0.99	56	1.8	2446
Qwen-1.5B	0.96	0.94	0.99	122	2.5	1121
Qwen-3B	0.96	0.95	0.99	182	2.8	751
Qwen-7B	0.96	0.95	0.99	308	4.4	443
Llama-1b	0.97	0.96	0.99	108	3	1271
Llama-3b	0.97	0.97	0.99	176	5.4	776
Llama-8b	0.97	0.97	0.99	322	6.2	423
Llama-70B*	0.75	0.97	0.98	12870	-	10
Qwen-72B*	0.80	0.97	0.98	16315	-	8
pyDeid	0.67	0.68	0.99	31	0.2	4490
obi-deid-bert	0.91	0.94	0.99	41	1.4	3333
Presidio	0.61	0.91	0.89	28	0.1	4931

Table 1: **Experiment 1:** Performance and GPU-based inference time for various de-identification models. We report the words-per-second rate, along with the time required to process 100 notes and the standard deviation across 10 runs to de-identify 100 notes. *Evaluated only once due to compute cost. **P: Precision, R: Recall**

5.2 Clinical Information Retention

Recent work (Aghakasiri et al., 2025) has demonstrated that standard classification metrics do not provide a complete picture of model performance. Specifically, when a false positive occurs, clinically relevant information may be removed. This reduces the utility of notes and thus negates the point of de-identification. Balancing the preservation of clinical utility with data sensitivity is critical. Therefore, in this work, we use the CIRE metric proposed by Aghakasiri et al. (2025), which uses an LLM to measure retention of clinical information by calculating the proportion of sentences that have changed clinically relevant information (CIRE prompt is presented in Appendix Figure 9).

5.3 Inference Efficiency

We also report the number of words/second. We prioritize this metric over the commonly references seconds/note, as the latter depends on the note length, which varies.

6 Experiments and Results

6.1 Experiment 1: LLM-based de-identification

Inference performance was measured on an NVIDIA A100-SXM4-80GB GPU. All models (except Llama-70B and Qwen-72B) were fine-tuned on 1,000 MIMIC-III discharge summaries. Llama-70B and Qwen-72B were prompt-tuned rather than fine-tuned on five discharge notes. For all fine-tuned models, inference time was computed on a test set of 100 discharge notes, repeated 10 times. The inference compute time of prompt-tuned models was measured only once due to the high inference time.

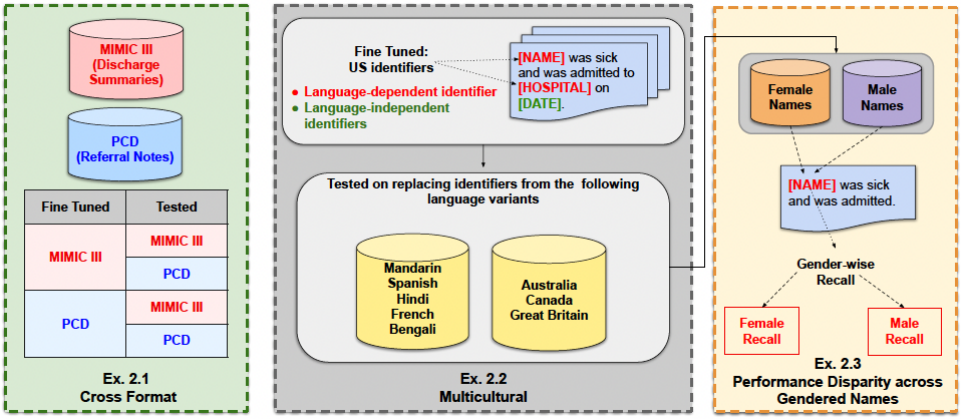

Figure 2: Setup of generalization testing. **(Ex. 2.1) Cross-format testing:** Models were fine-tuned on MIMIC and tested on both MIMIC and PCD, and vice versa, fine-tuned on PCD and tested on both datasets. **(Ex. 2.2) Multi-cultural testing:** Models fine-tuned on MIMIC notes with US English identifiers were tested on notes with identifiers from different language variants. **(Ex. 2.3) Performance disparity across gendered names:** Notes with identifiers from different language variants were used, and recall was evaluated specifically for name identifiers.

Table 1 highlights clear trade-offs between model size, performance, and computational efficiency. BERT, ModernBERT and ClinicalBERT achieved near-perfect precision and recall (0.99 each) with relatively low inference times ranging from 1528 to 2446 words/sec. For LLMs, the smaller variants (1B–8B) achieved high precision and recall (0.96–0.97) with inference ranging from 423 to 1271 words/sec. In contrast, Llama-70B and Qwen-72B processed 100 notes in more than 12,000 seconds (more than 3 hours), achieving only 8 to 10 words/sec, even when using two GPUs. Although they achieved high recall, their precision was not competitive, underscoring the limited practicality of large LLMs for de-identification. For subsequent experiments, we dropped Qwen2.5-72B, as its performance was comparable to Llama3.3-70B but was much slower during inference.

Rule-based systems exhibited the fastest runtimes, but at the expense of performance. Presidio achieved the fastest runtime (4931 words/sec) but had poor precision (0.61). Similarly, pyDeid was efficient (4490 words/sec), but underperformed in precision and recall (0.67 and 0.68). In contrast, obi-deid-bert (fine-tuned on i2b2) offered a more balanced trade-off, with moderate inference time (3333 words/sec) and higher recall (0.94). All models performed well in the CIRE score with no meaningful differences between model classes (except Presidio, which was about 0.10 lower than all other models).

Train - Test	PCD-PCD		PCD–M		M–M		M–PCD	
Model	P	R	P	R	P	R	P	R
BERT	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98
Qwen-7b	0.95	0.90	0.71	0.64	0.96	0.95	0.93	0.90
Llama-8b	0.98	0.97	0.93	0.85	0.98	0.97	0.96	0.94
Llama-70b*	0.92	0.98	0.70	0.97	0.72	0.97	0.93	0.97

Table 2: **Experiment 2.1 - Cross-format Evaluation:** Models fine-tuned on Private Clinical Dataset and MIMIC-III separately and tested on both datasets. *Prompt-tuned. **P: Precision, R: Recall, M: MIMIC**

6.2 Experiment 2: Generalization

We conducted three experiments (cross-format, multi-cultural and performance disparity across gendered names) to evaluate the robustness and generalization of de-identification models, illustrated in Figure 2.

Experiment 2.1: Cross-Format Generalization

To evaluate cross-format generalization, we fine-tuned the best performing models from Table 1 on 1,000 MIMIC-III discharge summaries and tested them on a separate collection of 1,000 MIMIC-III discharge notes (the MIMIC-test set) and 20% of the private data set of referral notes. In contrast, we also fine-tuned the models using 80% of the private dataset then tested them on the MIMIC-test set and 20% of the private dataset. This setup allowed us to assess robustness across different types of clinical notes and narrative styles. The results of this evaluation are shown in Table 2.

BERT achieved consistently high performance across all train–test combinations, with precision and recall near 0.99, indicating strong generalization across formats. Llama-8B also demonstrated

Model	Mandarin	Spanish	Hindi	French	Bengali
BERT	0.80	0.77	0.80	0.83	0.75
ModernBERT	0.75	0.83	0.92	0.90	0.81
ClinicalBERT	0.87	0.84	**0.97**	0.87	**0.97**
Qwen-1.5B	0.80	0.86	0.91	0.89	0.80
Qwen-3B	0.65	0.82	0.90	0.89	0.89
Qwen-7B	0.71	0.85	0.89	0.89	0.76
Llama-1b	0.83	0.80	0.90	0.87	0.77
Llama-3b	0.71	0.82	0.88	0.87	0.76
Llama-8b	0.76	0.85	0.91	0.90	0.78
Llama70	**0.96**	**0.95**	**0.97**	**0.97**	0.96
pyDeid	0.58	0.64	0.62	0.66	0.59
obi-deid-bert	0.94	0.84	0.92	0.90	0.93
Presidio	0.94	0.90	0.87	0.92	0.85

Figure 3: **Experiment 2.2.1:** Recall of de-identification models fine-tuned on 1000 MIMIC-III and tested on 500 samples from five languages. **P: Precision, R: Recall**. Full results in Appendix H.

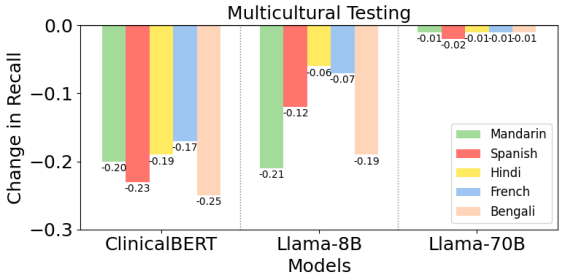

Figure 4: Relative difference in recall of the same model tested on US identifiers versus other languages. The models is fine-tuned on 1,000 samples with US English identifiers and evaluated on 500 samples for each language.

robust performance, though slightly lower in cross-format testing (PCD-M). This decline appears to be driven by variations in the narrative style between institutions. Surprisingly, Qwen-7B experienced a substantial drop in performance, nearly a 30-point drop evaluating on (PCD-M). Llama-70B, which was prompt tuned (with prompts for MIMIC-III and PCD provided in Figure 7 and Figure 8, respectively), maintained high recall across datasets, but tended to over-mask, resulting in reduced precision in some cross-dataset scenarios.

Experiment 2.2: Multi-Cultural Generalization

To evaluate how de-identification models trained on US English identifiers perform on identifiers from other language variants written in English (henceforth: multi-cultural generalization), instead of using the standard Faker settings, we replaced the identifiers using specialized settings in the Faker library. We picked the languages most spoken in the world (Mandarin, Spanish, Hindi, French, and Bengali)[1] (Ethnologue, 2025). Note: unlike previous work, our changes were more comprehensive and included changes beyond just names to other identifiers (e.g., addresses, phone numbers, and other identifier types).

This set of experiments used the MIMIC dataset. The models were fine-tuned on clinical notes with US English identifiers and then tested on notes with identifiers from other language variants.

Table 3 reports the recall of de-identification models trained in US-English and tested on identifiers from five non-English language variants (Mandarin, Spanish, Hindi, French, Bengali), with full results presented in Appendix Table 7. We found substantial recall drops for all models except for Llama-70B. More specifically, BERT, ModernBERT and ClinicalBERT maintained high precision, and ClinicalBERT achieved the highest recall for several language variants (Hindi and Bengali). Llama-1B and Llama-8B achieved strong recall, while prompt-tuned Llama-70B maintained very high recall but exhibited overmasking, leading to reduced precision.

Rule-based systems such as PyDeid and Presidio do not mask certain identifiers (e.g., hospital names, addresses, phone numbers) and were therefore evaluated differently. Only the identifiers these models actually mask were considered during the evaluation. For specific cases, the evaluation can be nondeterministic; for example, if a hospital name includes a personal name (e.g., 'Jack's Clinic'), the model may mask 'Jack' but not the full entity, which can lead to variations in reported metrics.

Appendix Table 8 presents the performance of models on identifiers from English variants (Great Britain, Australia, Canada). All fine-tuned transformer models maintained high precision and recall, demonstrating robustness to variations in addresses, phone numbers, and other identifiers in English-speaking regions. Prompt-tuned Llama-70B again showed high recall but decreased precision due to over-masking. These results indicate that, while models generalize well across English variations of identifiers, multi-lingual identifier variations remain more challenging, requiring careful model selection and tuning. Figure 4 presents the results for the best performing models for this set of experiments.

[1]Arabic, although among the top five languages, was excluded from the full evaluation due to limitations in the Faker library, which does not provide well-defined identifiers for all PII categories in Arabic.

Model	Mandarin		Spanish		Hindi		French		Bengali		GB		AU		CA	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT (all)	0.986	0.999	0.982	0.998	0.980	0.998	0.980	0.998	0.980	0.998	0.979	0.996	0.977	0.995	0.977	0.996
BERT (all–cult)	0.988	0.993	0.984	0.981	0.982	0.997	0.980	0.993	0.981	0.995	0.982	0.975	0.982	0.988	0.982	0.990
ModernBERT (all)	0.984	0.998	0.978	0.997	0.976	0.997	0.973	0.995	0.977	0.997	0.973	0.995	0.971	0.993	0.971	0.994
ModernBERT (all–cult)	0.985	0.996	0.978	0.984	0.974	0.998	0.979	0.991	0.975	0.976	0.971	0.976	0.977	0.992	0.975	0.989
ClinicalBERT (all)	0.980	0.998	0.975	0.999	0.973	0.998	0.973	0.998	0.974	0.997	0.971	0.995	0.970	0.994	0.970	0.995
ClinicalBERT (all–cult)	0.982	0.951	0.980	0.993	0.977	0.997	0.976	0.992	0.976	0.995	0.974	0.992	0.974	0.992	0.973	0.991

Table 3: **Experiment 3:** Performance of different de-identification models fine-tuned on MIMIC-III across language variants. "all" = trained on all variants; "all–cult" = trained on all variants except the target variant. **P: Precision, R: Recall**

Experiment 2.3: Performance disparity across Gendered Names

We evaluated the de-identification models, fine-tuned on MIMIC-III with US-based identifiers, using names drawn from the same set of languages as in Experiment 2.2. To assess potential gender bias, we separately measured performance on masculine and feminine names, allowing us to analyze model robustness across gendered name variations. Specifically, we computed the recall for each gender.

Appendix Table 9 summarizes the recall performance of de-identification models for feminine (R_f) and masculine (R_m) names across multiple languages (Mandarin, Spanish, Hindi, French, Bengali) and English variants (Great Britain). Almost all models achieved similar recall between genders, with differences generally below 0.05. Qwen variants exhibited a larger gap for French, where the difference in recall exceeded 0.05.

6.3 Experiment 3: Developing BERT-MultiCulture-DEID

As highlighted in Table 3, the obi-deid-bert model experienced substantial drops in model recall depending on the model evaluated (with up to a 10% drop in recall in Spanish). Unfortunately, this is the only readily accessible publicly available de-identification model accessible to most data curators. To address this gap, we sought to explore the feasibility of improving the performance of BERT-based de-identification on identifiers from unseen language variants.

For this experiment, we selected three BERT variants: BERT, ClinicalBERT, and ModernBERT (which can reasonably serve as replacements to the existing model). To perform this experiment, we fine-tuned two variations of each model. First, we trained models on all of the language variants present in Tables 3 and 8. Second, we trained models on all variants except the one being evaluated (e.g., the Mandarin evaluation would be trained on all other language variants except for Mandarin). The observed difference in performance serves as an indicator of the generalization gap exhibited by the model.

Table 3 presents the results of this experiment. First, we observe that training on all language variants improves performance on all variants (since these variants are no longer "out-of-distribution"). Surprisingly, we observe that for most models, the generalization gap is minimal (less than 1 percent). This indicates that the trained model, being exposed to a few language variants, became more robust.

7 Conclusion

In this study, we systematically evaluated the performance, generalization, and efficiency of various de-identification models across multiple scenarios, including across formats, language variants, and gendered names.

We observed that large LLMs are a degree of magnitude less efficient at de-identification compared to smaller LLM variants or BERT-based models. While large LLMs are generally more robust, more efficient models (e.g., BERT models) could be fine-tuned to better performance at very low cost, with only marginal improvements in performance increasing from 250 training notes to 1000 training notes. We found no disparity in model performance with respect to gendered names.

To improve the robustness of a popular and widely used publicly available BERT-based de-identification model (OBI Organization, 2022), we developed BERT-MultiCulture-DEID, a fine-tuned BERT model that demonstrates improved generalization to identifiers from multiple language variants.

7.1 Future Work

Our work has uncovered avenues for future work. The first and most direct step is to expand the set of language variants evaluated. Expanding the analysis to less-resourced languages is important to

ensure equitable protection of privacy. As newer models are released, their performance will need to be scrutinized in a similar way. We also need more publicly available benchmarking datasets to standardize and ensure transparency of future evaluations across studies.

We attempted to improve generalizability using an increased variety of data. Researchers can explore novel model architectures to improve the robustness of models without needing as much variety in the training data.

Limitations

We acknowledge that our study has limitations. First, inference was performed under controlled hardware specifications (e.g., 2 GPUs). The findings (especially the specific measurements) are likely to change with other setups (e.g. faster CPUs or GPUs would lead to different results), though we would expect relative performance to stay the same.

Second, while the Faker library simplifies the replacement of PII in clinical notes, it uses fixed distributions for names, addresses, dates, and other identifiers. This may not fully capture real-world, multi-language diversity, potentially introducing biases during fine-tuning. Additionally, there is a disparity in the amount of variation available to different Faker locales (i.e., some languages only have a small pool of identifiers or no identifiers for specific identifier categories), which affects the overall distributions.

Another limitation stemming from the Faker library deals with gendered languages. Our analysis relies on Faker's pre-defined list which only deals with the male and female gender and does not meaningfully deal with the issues surrounding name-based gender identification. While we are aware of these issues, we believe that it is still vital to attempt to uncover any performance discrepancies which may negatively affect patient privacy. For this reason, we proceeded with the analysis.

Moreover, our evaluation of multi-cultural generalization does not cover the entire spectrum of language and cultural variations. Although there is a great deal of linguistic variety in the evaluated languages, we have not proven that our results necessarily generalize to all languages.

Another limitation of our approach is that clinical language and practices evolve over time, which could lead to degradation of our fine-tuned models without periodic updates or retraining. Similarly, applying these models to different datasets may result in differences in performance and ranking.

Due to computational constraints (computational calculations are summarized in F), we were also limited in the complexity and number of experiments that we could perform. Specifically, fine-tuning smaller LLM variant models (e.g., 8B parameters) was limited to a maximum of 2,000 training samples. We also could not test all models for all subexperiments due to the exorbitant cost associated with training larger LLMs. Thus, we limited our experiments to specific models that were most likely to be of use to the research community.

A further limitation is the presence of labeling errors in the MIMIC dataset. Some elements of PII are incorrectly categorized (e.g., identifiers such as name or organization were mislabeled). Despite labeling errors, the tokens still represent PII and are usable for de-identification. Past work manually evaluating the accuracy of the MIMIC data set's PII label found few errors, concluding that it would have no meaningful impact on their results (Aghakasiri et al., 2025).

We believe that our work is generally low-risk as work serves to improve the robustness and efficiency of de-identification which in turn reduces the privacy risk of other works. We believe that the primary effect of our research is positive, though there are negative externalities with the execution of our work (e.g., the climate cost of running so many different models). Our work is limited in that we do not account for such externalities.

Acknowledgments

Carrie Ye is supported by a CRAF (CIORA)-Arthritis Society Canada New Clinical investigator Award (award #Cl-24-0013). Ross Mitchell is the Alberta Health Services Chair in Artificial Intelligence in Health and is supported by CIFAR, University Hospital Foundation, Amii, and the Canadian Foundation for Innovation. Mohamed Abdalla is supported by a CIFAR AI chair. Noopur Zambare is supported through an Amii grant.

References

Kiana Aghakasiri, Noopur Zambare, JoAnn Thai, Carrie Ye, Mayur Mehta, J Ross Mitchell, and Mohamed Abdalla. 2025. Not what the doctor ordered: Surveying llm-based de-identification and quantifying clinical information loss. *arXiv preprint arXiv:2509.14464*.

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

Bayan Altalla', Sameera Abdalla, Ahmad Altamimi, Layla Bitar, Amal Al Omari, Ramiz Kardan, and Iyad Sultan. 2025. Evaluating gpt models for clinical note de-identification. *Scientific Reports*, 15(1):3852.

Amazon Web Services. 2018. Amazon comprehend medical. https://aws.amazon.com/comprehend/medical/.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mikhail Budnikov, Anna Bykova, and Ivan P Yamshchikov. 2025. Generalization potential of large language models. *Neural Computing and Applications*, 37(4):1973–1997.

Centers for Medicare & Medicaid Services. 1996. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/.

Fangyi Chen, Syed Mohtashim Abbas Bokhari, Kenrick Cato, Gamze Gürsoy, and Sarah Rossetti. 2024. Examining the generalizability of pretrained de-identification transformer models on narrative nursing notes. *Applied Clinical Informatics*, 15(02):357–367.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, and 1 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Océane Dorémus, Dylan Russon, Benjamin Contrand, Ariel Guerra-Adames, Marta Avalos-Fernandez, Cédric Gil-Jardiné, Emmanuel Lagarde, and 1 others. 2025. Harnessing moderate-sized language models for reliable patient data deidentification in emergency department records: Algorithm development, validation, and implementation study. *JMIR AI*, 4(1):e57828.

David A Dorr, WF Phillips, Shobha Phansalkar, Shannon A Sims, and John Franklin Hurdle. 2006. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of information in medicine*, 45(03):246–252.

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

Ethnologue. 2025. What are the top 200 most spoken languages? https://www.ethnologue.com/insights/ethnologue200/.

European Union. 2016. General data protection regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

Daniele Faraglia and Other Contributors. 2014. Faker.

Óscar Ferrández, Brett R South, Shuying Shen, F Jeff Friedlin, Matthew H Samore, and Stéphane M Meystre. 2012. Generalizability and comparison of automatic clinical text de-identification methods and resources. In *AMIA Annual Symposium Proceedings*, volume 2012, page 199.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul M Heider, Jihad S Obeid, and Stéphane M Meystre. 2020. A comparative analysis of speed and accuracy for three off-the-shelf de-identification tools. *AMIA Summits on Translational Science Proceedings*, 2020:241.

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

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Alistair EW Johnson, Lucas Bulgarelli, and Tom J Pollard. 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.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Woojin Kim, Sungeun Hahm, and Jaejin Lee. 2024. Generalizing clinical de-identification models by privacy-safe data augmentation using gpt-4. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21204–21218.

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.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Stephane M Meystre, Gary Underwood, and Paul Heider. 2023. Clinideid, an open source solution for accurate clinical text de-identification. *Studies in Health Technology and Informatics*.

Microsoft. 2018. Presidio - Data Protection and De-identification SDK.

Mindee. 2021. doctr: Document text recognition. https://github.com/mindee/doctr.

Callandra Moore, Lucas Bulgarelli, Tom Pollard, and Alistair Johnson. 2023. Transformer-deid: Deidentification of free-text clinical notes with transformers.

Marco Naguib, Xavier Tannier, and Aurelie Neveol. 2024. Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852.

National Library of Medicine. 2019. Nlm scrubber. https://lhncbc.nlm.nih.gov/scrubber/.

Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B 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.

Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, and 1 others. 2020. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, 3(1):57.

OBI Organization. 2022. deid_bert_i2b2. https://github.com/obi-ml-public/ehr_deidentification.

David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André V Carreiro, and Vitor Rolla. 2024. Unlocking the potential of large language models for clinical text anonymization: A comparative study. *arXiv preprint arXiv:2406.00062*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training. *OpenAI technical report*.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Vaakesan Sundrelingam, Shireen Parimoo, Frances Pogacar, Radha Koppula, Saeha Shin, Chloe Pou-Prom, Surain B Roberts, Amol A Verma, and Fahad Razak. 2025. pydeid: an improved, fast, flexible, and generalizable rule-based approach for deidentification of free-text medical records. *JAMIA open*, 8(1):ooae152.

Latanya Sweeney. 1998. Strategies for de-identifying patient data for research.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Isabella C Wiest, Marie-Elisabeth Leßmann, Fabian Wolf, Dyke Ferber, Marko Van Treeck, Jiefu Zhu, Matthias P Ebert, Christoph Benedikt Westphalen, Martin Wermke, and Jakob Nikolas Kather. 2025. Deidentifying medical documents with local, privacy-preserving large language models: The llm-anonymizer. *NEJM AI*, 2(4):AIdbp2400537.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the name of fairness: assessing the bias in clinical record de-identification.

In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 123–137.

A Dataset Description

We sampled a total of 4,000 notes from MIMIC-III, which were divided into subsets for fine-tuning, validation, and evaluation. For fine-tuning, we experimented with training sets of varying sizes (250, 500, 1000, and 2000 notes) to study the effect of data scaling. The remaining notes were used for validation and final evaluation.

	PCD	MIMIC-III
Number of texts	204	4000
Note Length		
min	120	66
mean	670	1444
max	4975	6680
Number of Sensitive Words		
min	38	5
mean	183	70
max	981	362

Table 4: Descriptive statistics of datasets used in this paper.

B PII Entity Distribution

Figure 5 shows PII distribution in MIMIC-III test data.

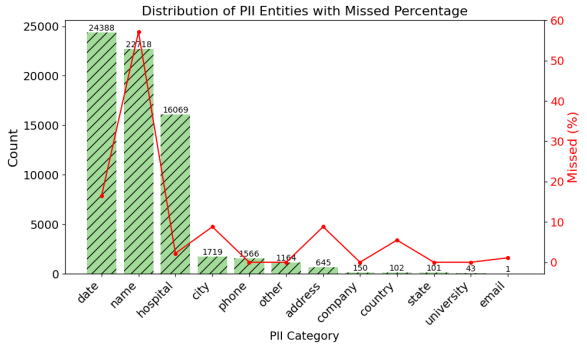

Figure 5: Distribution of PII entities in the testing data. The left y-axis shows the total count of each PII type in the test set, and the right y-axis shows the percentage of missed identifiers out of all missed identifiers during de-identification by the best-performing model BERT.

C Fine Tuning Setup

We used Python-based libraries, transformers (Wolf et al., 2020) and PEFT (Mangrulkar et al., 2022) for fine-tuning. BERT, ClinicalBERT, and ModernBERT were fine-tuned for 10 epochs with a learning rate of 2×10^{-5} and weight decay of

Model	Fine Tuning Samples	Precision	Recall
BERT	250	0.995	0.996
	500	0.997	0.998
	1000	0.997	0.998
	2000	0.998	0.999
ClinicalBERT	250	0.989	0.995
	500	0.992	0.996
	1000	0.996	0.998
	2000	0.997	0.998
ModernBERT	250	0.990	0.992
	500	0.994	0.996
	1000	0.996	0.998
	2000	0.998	0.998
Qwen 1.5B	250	0.93	0.91
	500	0.94	0.93
	1000	0.96	0.94
	2000	0.97	0.95
Qwen-3B	250	0.93	0.92
	500	0.95	0.94
	1000	0.96	0.95
	2000	0.97	0.96
Qwen-8B	250	0.92	0.92
	500	0.94	0.94
	1000	0.96	0.95
	2000	0.97	0.96
Llama-1B	250	0.96	0.94
	500	0.96	0.95
	1000	0.97	0.96
	2000	0.97	0.97
Llama-3B	250	0.95	0.95
	500	0.96	0.96
	1000	0.97	0.97
	2000	0.97	0.97
Llama-8B	250	0.95	0.96
	500	0.97	0.96
	1000	0.98	0.97
	2000	0.98	0.98

Table 5: Performance of various de-identification models fine-tuned on different numbers of MIMIC-III discharge summaries and evaluated on 1,000 MIMIC-III discharge summaries.

0.01. The batch size was 8 for BERT and ClinicalBERT, while it was 2 for ModernBERT. Due to their smaller context windows of 512, BERT and ClinicalBERT used a chunk size of 64, whereas ModernBERT, with a context window of 8192, used a chunk size of 1024.

All Llama and Qwen models were LoRA fine-tuned with $\alpha = 16$ and $r = 8$. Llama-1, Qwen-1.5, and Llama-3 and Qwen-3 were fine-tuned for 5 epochs. Llama-8B and Qwen-7B were trained for 3 epochs. All these models used a batch size of 1, chunk size of 2048, and the same learning rate (2×10^{-5}) and weight decay (0.01).

We also experimented with fine-tuning using different total numbers of samples, as summarized in Table 5.

D Multi-class Classification Results

We also fine-tuned BERT variants and smaller models from the Llama and Qwen families as multiclass classifiers to de-identify multiple types of PII (for e.g., names, phone numbers, addresses, countries, city, hospital name, company name, and other identifiers). They were equally effective in performance as their binary classifier counterparts.

Model	P	R
BERT	0.99	0.99
ModernBERT	0.99	0.99
ClinicalBERT	0.99	0.99
Qwen-1.5B	0.96	0.94
Qwen-3B	0.96	0.94
Qwen-7B	0.96	0.95
Llama-1B	0.97	0.95
Llama-3B	0.98	0.96
Llama-8B	0.98	0.97

Table 6: Performance of de-identification models fine-tuned as multiclass classifiers. **P: Precision, R: Recall**

E Prompts

Our prompt-tuning process was conducted using 5 clinical notes. It began with an initial prompt synthesized from prior work. From there, we conducted an iterative refinement cycle: after running the model on 5 sample notes, we performed qualitative error analysis to identify errors (e.g., missing categories). Based on these observations, we made modifications to the prompt, focusing on improving recall while preserving precision. After each iteration (prompts and results for some iterations are given in Figure 6, we re-evaluated performance (on the same 5 notes); we repeated this process until further adjustments no longer yielded measurable recall improvements. The final prompt was then evaluated on the test set, and the result is reported in the paper. Figure 7 presents the final prompt used for the MIMIC experiments with Llama-70B and Qwen-72B described in the paper. Figure 8 presents the prompt used for the experiments on the PCD dataset.

F Compute

The experiments described in the paper were performed on NVIDIA A100-SXM4-80GB GPUs. All fine-tuning and inference experiments (except for Llama-70B and Qwen-72B) were conducted on a single GPU. Llama-70B and Qwen-72B required two GPUs. Fine-tuning all models took approximately 44 hours in total (for all sample sizes 250, 500, 1000 and 2000). Testing all fine-tuned models, including baselines, required around 31 hours. Prompt tuning and testing of Llama-70B and Qwen-72B took approximately 255 hours, while calculating CIRE required about 252 hours.

The parameter specifications of the models were as follows: BERT and ClinicalBERT contained 110M parameters, while ModernBERT contained 149M parameters. The Llama family consisted of

Prompt 1 - Precision = 0.93, Recall = 0.98

You are a precise text transformer. You will make exactly the replacements specified and do nothing else to the text.
Please replace every instance of the following patient identifiers with their labels:
- ***Name***
- ***Phone_Number*** (even incomplete phone number)
- ***Address***
- ***Postal_Code***
- ***Medical_Record_Number(MRN)***
- ***Personal_Health_Number(PHN)***
- ***Unique_Lifetime_Identifier(ULI)***
- ***Date***
- ***Health_Card_Number(HCN)***
- ***Email***
- ***insurance_number***
- ***accession_number***
- ***provider_name***

IMPORTANT
- Do not truncate, summarize, or omit any part of the letter.
- Keep the sentence same if no identifier is present, do not add any comment or any extra text.
- Do not anonymize anything else (providers' info, lab values, normal text).
- All hospital names are also considered as address, and should be masked as [Address].
- Provider names includes name of doctor (or dr.), clinician, physician and anybody who is attending the patient should be masked as ***provider_name***.
- Return the entire sentence, verbatim except for the exact replacements above.
- Do not add anything extra.

Prompt 2 - Precision = 0.86, Recall = 0.99

You are a precise text transformer. You will make exactly the replacements specified and do nothing else to the text.
Please replace every instance of the following patient identifiers as well as all addresses, provider names and hospital/clinic names with their labels:
- [Name]
- [Phone_Number]
- [Address]
- [Postal_Code]
- [Medical_Record_Number(MRN)]
- [Personal_Health_Number(PHN)]
- [Unique_Lifetime_Identifier(ULI)]
- [Referral_Number]
- [Date]
- [Health_Card_Number(HCN)]
- [Patient_Email]
- [Insurance_Number]
- [Accession_Number]
- [Provider_Name]
- [Hospital]

Rules:
- Leave all other text completely untouched.
- Do not add any extra text or comments.
- Do not truncate, summarize, or omit any part of the sentence.
- Do not anonymize anything else (lab values, normal text).
- Provider names includes name of doctor (or dr.), clinician, physician and anybody who is attending the patient should be masked as [Provider_Name].
- If no identifier is present, return same sentence as input. Do not add any comment.
- Do not add 'The report does not contain any patient identifiers that need to be replaced' or 'This is the report'.
- Dates can be in any format (12/12/1999 or 12-1-2000 or 2001/09/02 or any any other format) and need to be masked as [Date]. Make sure no date is left unmasked. Do not mask time.
- Do not add any extra masking labels [medication, dosage, etc] apart from what is being said above.
- Mask all locality/places names, province names, country name as [Address].

Figure 6: Prompt-tuning Llama-70B with various prompts.

Prompt

You are a precise text transformer. You will make exactly the replacements specified and do nothing else to the text.
Please replace every instance of the following patient identifiers, as well as all addresses, provider names and hospital/clinic names with their labels:
- [Name]
- [Phone_Number]
- [Address]
- [Postal_Code]
- [Medical_Record_Number(MRN)]
- [Personal_Health_Number(PHN)]
- [Unique_Lifetime_Identifier(ULI)]
- [Referral_Number]
- [Date]
- [Health_Card_Number(HCN)]
- [Patient_Email]
- [Insurance_Number]
- [Accession_Number]
- [Provider_Name]
- [Hospital]

Rules:
- Leave all other text completely untouched.
- Do not add any extra text or comments.
- Do not truncate, summarise, or omit any part of the sentence.
- Do not anonymise anything else (lab values, normal text).
- Provider names include the name of the doctor (or dr.), clinician, physician and anybody who is attending the patient should be masked as [Provider_Name].
- If no identifier is present, return the same sentence as input. Do not add any comments.
- Dates can be in any format (12/12/1999 or 12-1-2000 or 2001/09/02 or any other format) and need to be masked as [Date]. Make sure no date is left unmasked. Do not mask time.
- Mask all locality/place names, province names, and country names as [Address].

Figure 7: Prompt used for de-identification using Llama-70B and Qwen-72B tuned on MIMIC dataset.

Prompt

You are a precise text transformer. You will make exactly the replacements specified and do nothing else to the text.
Please replace every instance of the following patient identifiers as well as all addresses, provider names and hospital/clinic names with their labels:
- [Name]
- [Phone_Number]
- [Address]
- [Medical_Record_Number(MRN)]
- [Personal_Health_Number(PHN)]
- [Unique_Lifetime_Identifier(ULI)]
- [Date]
- [Health_Card_Number(HCN)]
- [Email]
- [Insurance_Number]
- [Accession_Number]
- [Provider_Name]
- [Hospital]
- [Link]
- [Document_ID]
- [Other]

Rules:
- Leave all other text completely untouched.
- Do not add any extra text or comments.
- Do not truncate, summarize, or omit any part of the sentence.
- Do not anonymize anything else (lab values, normal text).
- Provider names includes name of doctor (or dr.), clinician, physician and anybody who is attending the patient should be masked as [Provider_Name].
- If no identifier is present, return same sentence as input. Do not add any comment. Do not add 'The report does not contain any patient identifiers that need to be replaced.
- Dates can be in any format (12/12/1999 or 12-1-2000 or 2001/09/02 or 2002-09-09 or 22Sep22 or any any other format) and need to be masked as [Date]. Do not mask time.
- Any weblink should be masked as [Link].
- Mask id in 'Printed by:' as [Other].
- Mask all locality/place names as [Address].

Figure 8: Prompt used for de-identification using Llama-70B and Qwen-72B tuned on PCD.

Prompt
You are given two sentences, labelled "original:" and "deid:". Your job is to decide only whether the de-identified sentence has altered any clinically meaningful information. Answer with Yes or No (nothing else).

Clinically meaningful changes include:
- Adding, removing, or modifying a medication, diagnosis, procedure, test result, clinical instruction, or patient history.
- Altering the patient's age.
- Any change that would affect medical interpretation or decision-making.

Do NOT count as clinical changes:
- Removal or masking of facility names, street addresses, clinic names, hospital names, or other location identifiers.
- Removal or masking of personal or institutional identifiers, such as patient or provider names, practice IDs, PHNs, MRNs, account numbers, etc.
- Generic role titles (e.g. nurse, physician) or department names.
- Removal of DOB (date of birth), years and dates.

Examples
Original: Prescribed ferric maltol capsules
deid: Prescribed maltol capsules
Output: Yes (because a clinically important word is deleted)

Original: The patient is Jerry, seen at East Clinic on 2025-04-01.
deid: The patient is [NAME], seen at [CLINIC] on [DATE].
Output: No (patient name and dates are not clinically relevant)

Original: Was tested positive on 13/05/2001, and had fever on 13th April 2013
deid: Was tested positive on [DATE], and had fever on [DATE]
Output: No (because this date refers to a date of testing)

Original: The patient is John, aged 39 has fever.
deid: The patient is [NAME], aged [AGE] has fever
Output: Yes (because AGE is removed)

Original: Patient MRN 123456 underwent colonoscopy on 2025-02-20 3.5 2 mg acetaminophen and was given anti-viral medicines.
deid: Patient MRN [MRN] underwent colonoscopy on [DATE][DATE][DATE] 3.5 2 mg acetaminophen and was given anti-viral medicines.
Output: No (MRN and dates are not clinically relevant and medications are preserved)

Please apply this to each pair and return only Yes or No.

Figure 9: Prompt used for CIRE with Llama-70B.

models with 1B, 3B, 8B, and 70B parameters, and the Qwen family included models with 1.5B, 3B, 7B, and 72B parameters.

G Use of AI Assistants

An AI assistant was used only for spelling, grammar, and phrasing.

H Full Results

Model	GB		AU		CA	
	P	R	P	R	P	R
BERT	0.99	0.97	0.99	0.98	0.99	0.99
ModernBERT	0.99	0.97	0.99	0.99	0.99	0.99
ClinicalBERT	0.99	0.96	0.99	0.99	0.99	0.99
Qwen-1.5B	0.96	0.93	0.96	0.95	0.96	0.94
Qwen-3B	0.97	0.94	0.96	0.95	0.96	0.95
Qwen-7B	0.97	0.94	0.97	0.95	0.96	0.95
Llama-1b	0.97	0.95	0.97	0.96	0.97	0.96
Llama-3b	0.97	0.96	0.97	0.97	0.97	0.96
Llama-8b	0.98	0.96	0.98	0.97	0.98	0.96
Llama-70B	0.75	0.98	0.74	0.98	0.76	0.97
pyDeid	0.68	0.71	0.67	0.72	0.68	0.74
obi-deid-bert	0.91	0.91	0.91	0.94	0.91	0.93
Presidio	0.63	0.88	0.62	0.92	0.62	0.92

Table 8: **Experiment 2.2.2:** Performance of de-identification models fine-tuned on 1000 MIMIC-III and evaluated on 500 samples from above 3 English variants. **P: Precision, R: Recall**

Model	Mandarin		Spanish		Hindi		French		Bengali	
	P	R	P	R	P	R	P	R	P	R
BERT	0.99	0.80	0.99	0.77	0.99	0.80	0.99	0.83	0.99	0.75
ModernBERT	0.99	0.75	0.99	0.83	0.99	0.92	0.99	0.90	0.99	0.81
ClinicalBERT	0.99	0.87	0.99	0.84	0.99	**0.97**	0.99	0.87	0.99	**0.97**
Qwen-1.5B	0.97	0.80	0.96	0.86	0.96	0.91	0.96	0.89	0.96	0.80
Qwen-3B	0.97	0.65	0.96	0.82	0.97	0.90	0.97	0.89	0.97	0.89
Qwen-7B	0.98	0.71	0.97	0.85	0.97	0.89	0.97	0.89	0.97	0.76
Llama-1b	0.98	0.83	0.97	0.80	0.97	0.90	0.97	0.87	0.97	0.77
Llama-3b	0.98	0.71	0.98	0.82	0.98	0.88	0.98	0.87	0.97	0.76
Llama-8b	0.99	0.76	0.98	0.85	0.98	0.91	0.98	0.90	0.98	0.78
Llama70-prompting	0.81	**0.96**	0.79	**0.95**	0.62	**0.97**	0.74	**0.97**	0.78	0.96
pyDeid	0.70	0.58	0.70	0.64	0.64	0.62	0.67	0.66	0.62	0.59
obi-deid-bert	0.94	0.94	0.92	0.84	0.91	0.92	0.91	0.90	0.91	0.93
Presidio	0.44	0.94	0.64	0.90	0.60	0.87	0.64	0.92	0.58	0.85

Table 7: **Experiment 2.2.1:** Performance of de-identification models fine-tuned on 1000 MIMIC-III and tested on 500 samples from five languages. **P: Precision, R: Recall**

Model	Mandarin		Spanish		Hindi		French		Bengali		GB	
	R_f	R_m	R_f	R_m	R_f	R_m	R_f	R_m	R_f	R_m	R_f	R_m
BERT	0.92	0.90	0.88	0.89	0.58	0.60	**0.90**	**0.96**	0.53	0.52	1.0	1.0
ModernBERT	0.85	0.86	0.94	0.93	0.82	0.80	0.96	0.98	0.76	0.76	0.99	0.99
ClinicalBERT	0.85	0.85	0.96	0.95	0.93	0.95	0.96	0.99	0.94	0.97	0.99	0.99
Qwen 1.5	0.92	0.91	0.81	0.83	0.78	0.82	**0.81**	**0.87**	0.70	0.72	**0.90**	**0.95**
Qwen 3	0.81	0.83	0.80	0.83	0.80	0.83	**0.83**	**0.88**	**0.71**	**0.76**	0.92	0.96
Qwen 7	0.87	0.84	0.82	0.84	0.72	0.75	**0.83**	**0.88**	0.62	0.64	0.92	0.95
Llama-1b	0.93	0.90	0.78	0.81	0.72	0.77	**0.82**	**0.88**	0.68	0.71	**0.90**	**0.95**
Llama-3b	0.86	0.85	0.81	0.83	0.70	0.72	**0.85**	**0.90**	0.65	0.70	**0.91**	**0.96**
Llama-8b	0.93	0.90	0.86	0.87	0.79	0.82	0.87	0.91	0.74	0.79	0.92	0.96
Llama-70B	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
pyDeid	**0.80**	**0.69**	0.78	0.75	0.63	0.63	0.72	0.70	0.63	0.62	**0.85**	**0.90**
obi-deid-bert	0.98	0.97	0.91	0.91	0.91	0.94	0.95	0.98	0.93	0.94	0.99	0.99
Presidio	0.97	0.96	0.88	0.90	0.84	0.84	0.93	0.95	**0.72**	**0.80**	0.97	0.97

Table 9: **Experiment 2.3:** Performance disparity across gendered names of de-identification models fine-tuned on 1000 MIMIC-III and tested on 500 samples having identifiers from six language variants. Differences in recall greater than 0.05 between feminine and masculine names are highlighted. The Faker library does not provide an explicit pool for female and male names for Australian and Canadian English. R_f**: Female Recall,** R_m**: Male Recall**