EdgeDeID: Advancing De-identification with Small LMs (SLMs) and Synthetic Data on Edge Devices

Bardia Khosravi, MD, MPH, MHPE, Radiology Resident, Department of Radiology, Yale University
Theo Dapamede, MD, PhD; Frank Li, PhD; Pouria Rouzrokh, MD, MPH, MHPE; Aawez Mansuri, MS; Elham Mahmoudi, MD, MPH; Rohan Satya Isaac, MS; Amirali Khosravi, MD; Mohammadreza Chavoshi, MD; Hari Trivedi, MD; Janice Newsome, MD; Bradley Erickson, MD, PhD, CII

Introduction

De-identification of medical text reports is crucial for maintaining patient privacy while enabling data sharing for research and analysis. Existing tools, primarily based on BERT, face two key challenges: (1) they process the entire text sequentially, leading to suboptimal performance when dealing with long reports containing minimal protected health information (PHI), and (2) they require significant computational resources, making them impractical for edge devices. This study presents EdgeDeID, a novel approach leveraging small language models (SLMs) and synthetic data for efficient de-identification on edge devices.

Hypothesis

We hypothesize that by utilizing a small language model fine-tuned on synthetically generated data, EdgeDeID can accurately extract PHI entities, significantly reducing processing time compared to BERT-based methods, especially on edge devices with limited resources.

Methods

Synthetic data was generated using the Hermes model, a fine-tuned LLaMA 405B, and augmented with techniques such as error introduction and name variations to create a diverse training set. The original training set contained 18,000 samples, which was further expanded to 23,000 through augmentations. The Qwen 2.5 Coder decoder-only transformer (0.5B parameters) was fine-tuned on this dataset using supervised learning. The model supports a context length of 8k tokens, surpassing other models limited to 512 tokens. Temperature zero was used for all inference cases. EdgeDeID's performance was evaluated on 100 manually annotated reports.

Results

EdgeDeID processed 100 reports in under 20 seconds on a GPU with 8GB RAM and 4.2 seconds per report on devices without a GPU. The model achieved an overall recall (sensitivity) of 98.47% across all entities, with the lowest recall of 95.0% for date-type entities, which can be further optimized using rule-based methods to ensure complete anonymity. Model's precision was 99.0%. Additionally, the model achieved >95% sensitivity across the 21PHI entities.

Conclusion

EdgeDeID presents a feasible solution for de-identifying medical reports on edge devices, addressing the challenges of data privacy when using LLMs that require external data transmission. By leveraging SLMs and synthetic data, EdgeDeID achieves rapid and accurate de-identification, facilitating secure data sharing for research and analysis.

Figure(s)

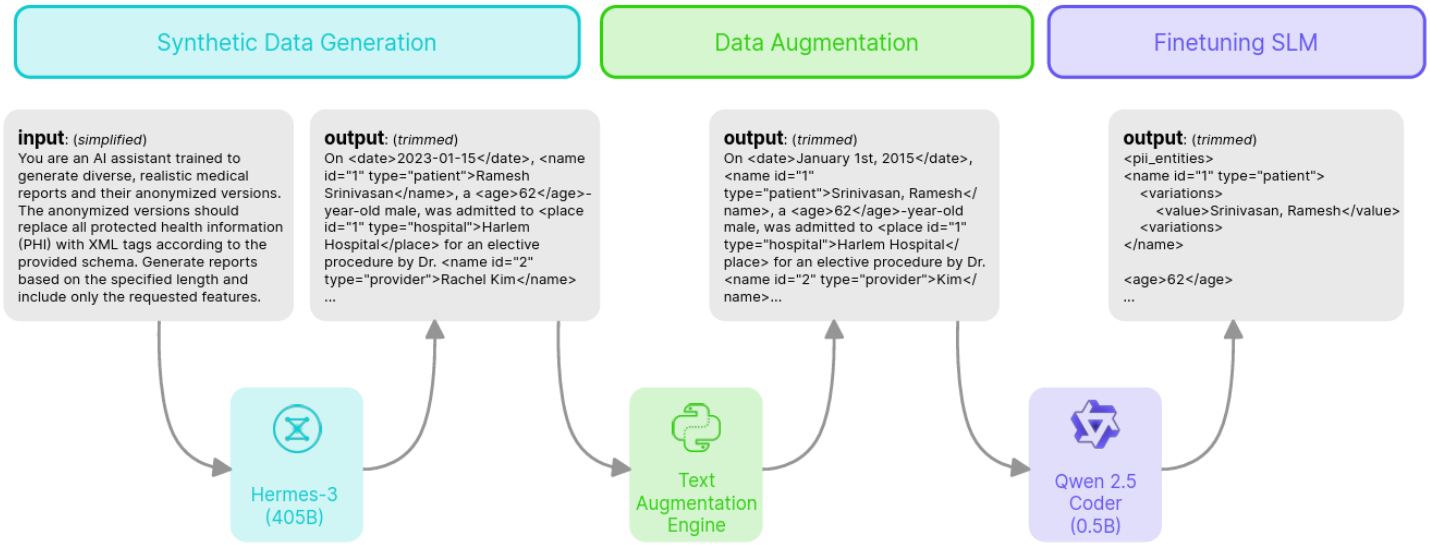

Figure 1. EdgeDeID Pipeline for Medical Text De-identification. The pipeline consists of three main stages: 1 Synthetic Data Generation using Hermes-3 405B parameters), which generates diverse medical reports with XML-annotated PHI entities based on specified inputs; 2 Data Augmentation through a custom Text Augmentation Engine that introduces controlled variations such as typos, formatting differences, and name variations to improve model robustness; and 3 Fine-tuning of Qwen 2.5 Coder 0.5B parameters), a small language model optimized for edge devices. The process converts unstructured medical text into a structured format which enables seamless programmatic manipulation for de-identification tasks, such as redaction or pseudonymization of sensitive information while maintaining the clinical narrative's integrity.

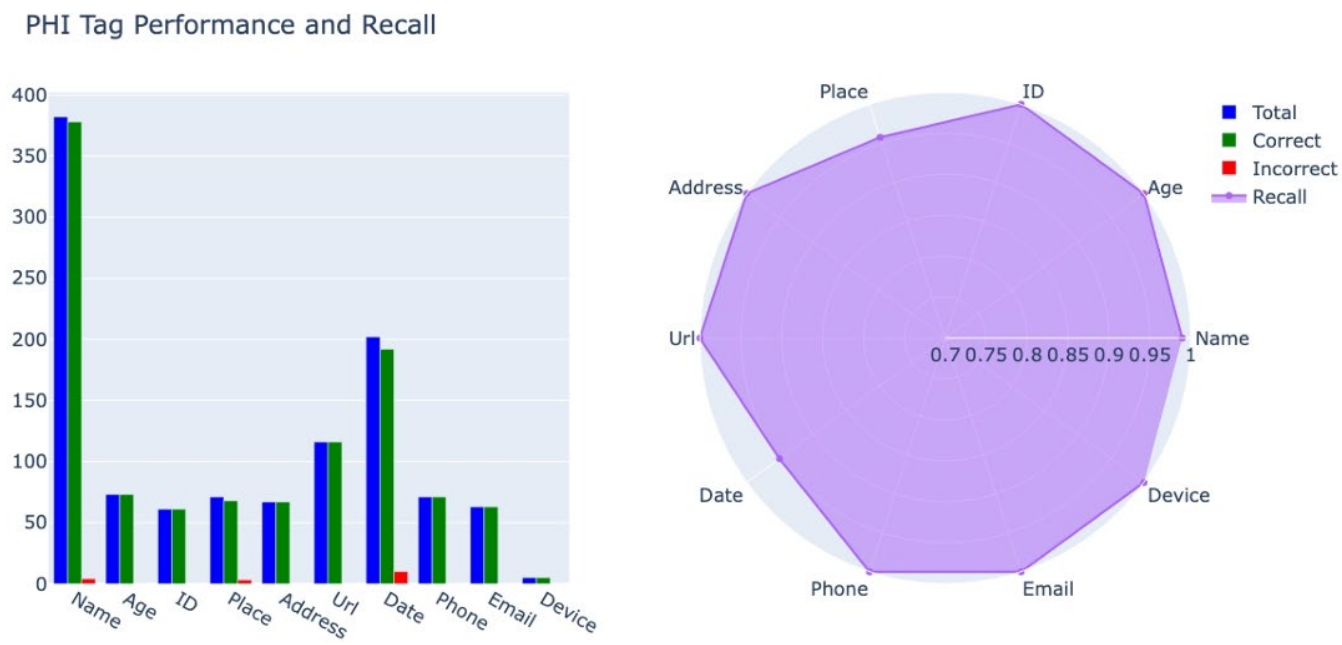

Figure 2. This figure shows the performance of EdgeDeID on the manually annotated test set for various Protected Health Information (PHI) tags. The left bar chart shows the total number of each tag, the number of correctly identified tags, and the number of incorrectly identified tags. The right radar chart shows the recall for each tag. The recall ranges from 0.95 to 1.0 for all tags.

Keywords

Applications; Artificial Intelligence/Machine Learning; Emerging Technologies; Security