Large Language Model Empowered Privacy-Protected Framework for PHI Annotation in Clinical Notes

Guanchen WU[a], Linzhi ZHENG[b], Han XIE[a], Zhen XIANG[c], Jiaying LU[d], Darren LIU[d], Delgersuren BOLD[d], Bo LI[b], Xiao HU[d], and Carl YANG[a,1]
[a] *Department of Computer Science, Emory University*
[b] *Department of Computer Science, University of Chicago*
[c] *School of Computing, University of Georgia*
[d] Nell Hodgson Woodruff School of Nursing, Emory University

Abstract. De-identifying private information in medical records is crucial to prevent confidentiality breaches. Rule-based and learning-based methods struggle with generalizability and require large annotated datasets, while LLMs offer better language comprehension but face privacy risks and high computational costs. We propose LPPA, an LLM-empowered Privacy-protected PHI Annotation framework, which uses few-shot learning with pre-trained LLMs to generate synthetic clinical notes, reducing the need for extensive datasets. By fine-tuning LLMs locally with synthetic notes, LPPA ensures strong privacy protection and high PHI annotation accuracy. Experiments confirm its effectiveness, efficiency, and scalability.

Keywords. LLM, clinical note, PHI, PHI annotation, de-identification

1. Introduction

Clinical notes, including discharge notes and nursing notes, are unstructured texts that document patient care[1]. Discharge notes, in particular, provide comprehensive accounts of hospital stays, covering physician observations, social determinants of health, and nuanced clinical details beyond structured EHR data. These notes are critical for understanding patient conditions and advancing care. However, sharing them for research is challenging due to the presence of sensitive Protected Health Information (PHI)[2], such as names and birth dates. To comply with HIPAA[3], PHI must be removed, complicating large-scale sharing of high-utility clinical data.

Existing methods for PHI de-identification include rule-based systems, learning-based approaches, and large language models (LLMs). Rule-based systems rely on predefined patterns and domain-specific dictionaries[4] but struggle with unstructured clinical narratives, making them inconsistent and difficult to scale. Learning-based methods, such as SVMs and CRFs, improve upon rule-based systems by learning features from annotated data but require extensive manual feature engineering and face challenges with domain generalization. Deep learning models like RNNs and LSTMs[6,7] reduce manual effort through automatic feature extraction but demand large datasets and

[1] Corresponding Author: Carl Yang, j.carlyang@emory.edu.

computational resources, limiting scalability. Recent advancements in pre-trained language models (PLMs), such as BERT, and LLMs, like GPT[8,9], improve contextual understanding and enable few-shot learning, reducing dependence on annotated datasets[10]. However, smaller LLMs lack accuracy, while larger models require high computational resources for local deployment. Using public APIs to mitigate these costs risks data exposure and non-compliance with privacy regulations like HIPAA.

In this work, we introduce LPPA, an LLM-empowered Privacy-protected PHI Annotation framework, to address these limitations. LPPA contributes by: 1) leveraging pre-trained LLMs and few-shot learning to generate synthetic clinical notes, removing the need for extensive manual work and large annotated datasets while preserving real-world language complexity; and 2) ensuring data privacy by fine-tuning a locally hosted LLM, eliminating reliance on external APIs. Our method achieves an F1 score of 0.57 on the real-world clinical note dataset, closely approaching the performance of state-of-the-art LLMs. This result underscores the framework's comparable accuracy while offering greater efficiency, scalability, and almost-zero reliance on annotated datasets, making it a practical and privacy-conscious solution for PHI annotation.

2. Method

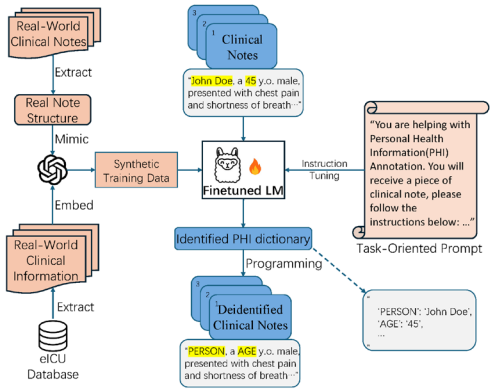

Figure 1. LPPA Framework Overview.

Our objective is to accurately detect and annotate PHI entities within clinical notes and replace them with corresponding entity types while minimizing the need for large amounts of manually annotated training data. As shown in Figure 1, the LPPA framework combines two synthetic data generation approaches to address data scarcity and ensure privacy. The first approach uses few-shot prompting with real-world clinical notes to generate synthetic data that mimics authentic documentation. The second approach employs anonymized public datasets, using simulated identifiers (e.g., names and addresses) to create realistic synthetic notes. These two datasets are mixed to mitigate quality bias, ensuring robust training. A locally fine-tuned LLM eliminates reliance on external APIs, reduces the need for manual feature engineering, and ensures high accuracy and privacy compliance.

2.1. LLM-based Synthetic Data Generation

Due to strict privacy regulations, access to real-world clinical notes is limited, making synthetic data generation using LLMs essential for augmenting training data in our framework. Although we possess a small dataset of fully annotated real notes, it is insufficient to train a robust model, and publicly available datasets lack the richness of real-world PHI. By leveraging LLMs, we generate realistic, high-quality synthetic notes that closely resemble real-world notes while adhering to privacy regulations. Our approach uses two approaches to generate synthetic notes that are structurally and

contextually aligned with real-world notes. By harnessing the language generation capabilities of LLMs, we generate synthetic clinical notes that incorporate diverse PHI entities, effectively mimicking the variability present in real-world clinical data.

Anonymized Example-Guided Note Generation. The first approach uses a few-shot prompting technique, where the LLM is guided by a few fully anonymized representative real-world clinical notes to learn the structure of authentic notes. These examples include key sections like patient demographics, medical history, diagnoses, and treatments. Using this template, the LLM is able to generate synthetic notes with embedded simulated PHI entities, ensuring the data replicates real-world structure and content while adhering to strict privacy standards.

Synthetic PHI Insertion into Anonymized Notes. The second approach leverages a public dataset to extract real-world clinical information, such as patient gender, allergies, diagnoses, medications, and lab results. This information serves as a foundation for the LLM to generate synthetic clinical notes. The LLM is tasked with creating simulated PHI identifiers, including plausible names, phone numbers, and addresses. These simulated identifiers are then integrated with the extracted clinical information to produce synthetic notes that closely mimic authentic medical records. This approach ensures the notes retain the richness and structure of real-world clinical documentation while maintaining strict compliance with privacy.

Synthetic Data Mixture. To address potential variations in the quality of synthetic training data generated by the two approaches, we randomly mixed the datasets. This ensured unbiased and consistent model performance, minimizing the impact of any data quality differences.

2.2. Instruction Tuning

To ensure comprehensive PHI extraction, the fine-tuning process uses prompts designed to identify all potential PHI entities, prioritizing recall to minimize missed sensitive information and comply with HIPAA. This approach achieves reliable de-identification, balancing security and usability in clinical settings.

3. Results

This study utilized 100 fully annotated real-world clinical notes from Emory Hospitals, carefully reviewed and scrambled for PHI integrity. Usage of the data has been approved under IRB number STUDY00006871. Due to privacy limitations, these notes were reserved for evaluation only. Additionally, the eICU Collaborative Research Database, a publicly available dataset containing de-identified records from over 200,000 ICU admissions, was used to extract clinical information for synthetic data generation.

We conduct experiments on 100 real-world clinical notes. The base model for the instruction-tuning process is the Llama-3-8B-Instrut model. Baselines included a rule-based method using regular expressions and several LLMs: Meta's Llama-3-8B and 70B models, as well as OpenAI's GPT models accessed via Azure to ensure data privacy. Performance was assessed using precision, recall, and F1-score, providing a comprehensive evaluation of PHI annotation accuracy, completeness, and overall effectiveness.

Table 1. This table presents the average evaluation results for different baseline models and our fine-tuned models on 100 real-world clinical notes. The “1K AEG-Tuned Model” was fine-tuned using 1,000 synthetic

notes generated through the **A**nonymized **E**xample-**G**uided Note Generation approach. The "SPI-Tuned Model" was fine-tuned with synthetic data generated via the **S**ynthetic **P**HI **I**nsertion approach. The "Hybrid-Tuned Model" integrates both tuning methods. The reported scores represent averages across all 100 clinical notes, and the "/" sign indicates that the model cannot identify this PHI category.

Models	**Overall**			**PERSON**			**AGE**			**DATE/TIME**		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
PhysioNet	0.39	0.24	0.28	0.01	0.20	0.02	/	/	/	0.91	0.76	0.83
Llama3-8B-Instruct	0.46	0.59	0.50	0.53	0.55	0.53	0.38	0.43	0.41	0.79	0.39	0.52
Llama3-70B-Instruct	0.60	0.68	0.62	0.59	0.53	0.56	0.48	0.42	0.45	0.83	0.50	0.63
GPT-3.5-turbo	0.43	0.60	0.48	0.60	0.50	0.54	0.39	0.36	0.37	0.74	0.44	0.55
GPT-4	0.53	0.69	0.58	0.60	0.57	0.58	0.45	0.37	0.41	0.80	0.57	0.67
1k AEG Model	0.47	0.57	0.50	0.48	0.52	0.50	0.49	0.45	0.47	0.66	0.43	0.52
2k AEG Model	0.54	0.61	0.55	0.54	0.53	0.54	0.50	0.44	0.46	0.71	0.44	0.54
3k AEG Model	0.55	0.61	0.56	0.56	0.54	0.55	0.51	0.41	0.45	0.73	0.42	0.53
1k SPI Model	0.51	0.54	0.50	0.58	0.54	0.56	0.48	0.42	0.45	0.72	0.40	0.51
2k SPI Model	0.52	0.53	0.50	0.55	0.50	0.52	0.52	0.41	0.46	0.49	0.34	0.48
3k Hybrid Model	0.59	0.53	0.53	0.61	0.54	0.57	0.51	0.44	0.47	0.76	0.36	0.49
4k Hybrid Model	0.65	0.54	0.57	0.59	0.53	0.56	0.49	0.42	0.45	0.82	0.40	0.54
5k Hybrid Model	0.64	0.55	0.57	0.59	0.53	0.56	0.51	0.44	0.48	0.82	0.35	0.50

Table 1 evaluates the performance of baseline models and fine-tuned models on real-world clinical notes, highlighting the effectiveness of different tuning approaches for PHI annotation. Baseline models, while demonstrating moderate general performance, struggle with certain entity types, particularly PERSON and AGE, whereas structured entities like DATE/TIME are more effectively identified. Fine-tuned models, especially those utilizing hybrid strategies that combine data from two synthetic data generation approaches, show notable improvements in overall performance. These results emphasize the superiority of hybrid-tuned models in achieving balanced and robust performance across diverse PHI categories.

4. Discussion

The results emphasize the importance of balancing data size during fine-tuning to manage trade-offs between precision and recall in PHI annotation. Fine-tuned models trained on larger datasets show notable improvements, particularly for the hybrid-tuned approach, which effectively enhances generalization. However, trade-offs persist, as precision and recall vary across entity types, with unstructured categories like PERSON posing greater challenges compared to structured ones like DATE/TIME. These findings suggest that optimizing synthetic data generation to address specific entity characteristics while ensuring balanced representation is critical for improving model robustness and adaptability in real-world applications.

While our study evaluates the utility of both commercial and open-source LLMs for de-identifying PHI, we acknowledge the absence of classical natural language processing

(NLP) approaches, such as named entity recognition (NER) models, in our comparative analysis. Furthermore, the inclusion of previously established PHI de-identification benchmarks would have provided a valuable complementary perspective alongside our evaluation using private clinical notes.

5. Conclusions

We proposed LPPA framework, leveraging LLMs for PHI de-identification in clinical notes. By combining synthetic data generation and instruction tuning, LPPA reduces the need for large annotated datasets while maintaining high recall and precision. Our fine-tuned models demonstrate competitive accuracy with minimal manual annotation, offering a scalable, privacy-conscious, and efficient solution for real-world clinical use. Beyond healthcare, LPPA framework has potential applications in domains like legal and financial document analysis where privacy-preserving text processing is critical. Future work could enhance the framework by incorporating multimodal data, improving model generalizability for diverse PHI types, and optimizing the trade-off between computational cost and performance. Adapting the framework to comply with stricter privacy regulations would further enable its adoption across diverse industries.

Acknowledgements

This research was partially supported by the US National Science Foundation under Award Number 2319449 and Award Number 2312502, as well as the US National Institute of Diabetes and Digestive and Kidney Diseases of the US National Institutes of Health under Award Number K25DK135913. This research project has benefited from the Microsoft Accelerating Foundation Models Research (AFMR) grant program.

References

[1] Boag W, Doss D, Naumann T, Szolovits P. What's in a note? unpacking predictive value in clinical note representations. AMIA Summits on Translational Science Proceedings. 2018 May 18;2018:26.
[2] Moore W, Frye S. Review of HIPAA, part 1: history, protected health information, and privacy and security rules. Journal of nuclear medicine technology. 2019 Dec 1;47(4):269-72.
[3] Cohen IG, Mello MM. HIPAA and protecting health information in the 21st century. Jama. 2018 Jul 17;320(3):231-2.
[4] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. American journal of clinical pathology. 2004 Feb 1;121(2):176-86.
[5] Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017 May;24(3):596-606.
[6] Khin K, Burckhardt P, Padman R. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. arXiv preprint arXiv:1810.01570. 2018 Oct 3.
[7] Liu Z, Huang Y, Yu X, Zhang L, Wu Z, Cao C, Dai H, Zhao L, Li Y, Shu P, Zeng F. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032. 2023 Mar 20.
[8] Yashwanth YS, Shettar R. Zero and few short learning using large language models for de-identification of medical records. IEEE Access. 2024 Aug 7.
[9] Wu G, Ling C, Graetz I, Zhao L. Ontology extension by online clustering with large language model agents. Frontiers in Big Data. 2024 Oct 7;7:1463543.