Design and Evaluation of a Post-NER Refinement Pipeline for Clinical Note De-identification

Zejun Zhou
Brown University
zejun_zhou@brown.edu

***Abstract*—Clinical note de-identification is critical for protecting patient privacy and enabling the safe use of medical text in research and machine learning applications. However, existing named entity recognition (NER)-based approaches often struggle under real-world variation and typically provide limited support for auditing and error analysis. By examining the outputs of NER-based systems, this project identifies three recurring failure modes in clinical de-identification pipelines: false positives, false negatives, and fragmented entity spans, where fragmented spans can simultaneously introduce both error types under exact-match evaluation.**

This project presents a transparent, locally deployable post-NER de-identification framework that augments NER-based extraction with two refinement stages: a Verification loop for correcting candidate entities and a Candidate Expansion loop for recovering missed protected health information (PHI). Beyond improving extraction quality, the system incorporates transparency-oriented components, including structured error attribution, audit-ready outputs, and analytical tooling for human review and iterative debugging.

The framework is evaluated on the i2b2 2014 benchmark, a MIMIC-IV radiology subset, and a synthetic dataset designed to simulate distribution shift. Experimental results show substantial robustness improvements under variation, increasing entity-level F1 from 61.52% to 87.78% on the synthetic dataset while significantly reducing false negatives. These results suggest that combining post-NER refinement with transparency-oriented infrastructure can improve both the robustness and inspectability of clinical de-identification pipelines.

I. Introduction

Clinical notes contain rich information about patient conditions, treatments, and clinical decision-making, making them valuable for medical research and machine learning applications. However, these records often include sensitive information, such as names, dates, locations, and identifiers, which must be removed prior to data sharing. De-identification systems aim to automatically detect and remove such protected health information (PHI) from clinical text, serving as a critical enabler for privacy-preserving data use.

Many existing approaches formulate de-identification as a sequence labeling task and apply named entity recognition (NER) models to identify PHI spans. While modern NER models achieve strong results on benchmark datasets, their performance often degrades when applied to clinical notes that differ in writing style, annotation conventions, or domain-specific terminology. In real-world clinical settings, such variability is common, raising concerns about the robustness and reliability of these systems. In addition, many de-identification pipelines operate largely as black boxes, providing limited support for understanding, auditing, or correcting extraction errors, capabilities that are important in clinical and compliance-oriented workflows.

To better understand these limitations, this project analyzes the behavior of a widely used de-identification model across datasets with distinct characteristics, including the i2b2 2014 benchmark dataset [3], a radiology subset derived from MIMIC-IV [8], and a synthetic dataset designed to simulate distribution shift. The analysis shows that the same model can exhibit substantially different behavior across datasets, indicating that extraction quality is highly sensitive to dataset variation. A closer inspection of model outputs reveals three recurring error patterns: false positives, false negatives, and fragmented entity spans. Fragmented spans are particularly problematic under strict span-level evaluation because they can simultaneously introduce both false positives and false negatives.

To address these challenges, this project develops a *transparent, locally deployable de-identification framework* that refines NER predictions through two complementary stages: (i) a Verification loop and (ii) a Candidate Expansion loop. Rather than re-extracting entities from scratch, the framework treats the output of the NER model as an initial set of candidate spans and applies targeted refinement to improve extraction quality.

The Verification loop examines each predicted span in the context of the full clinical note and determines whether the span should be retained, removed, or adjusted. This stage addresses two common failure modes of NER systems: false positives and fragmented entity spans. In particular, the verification process performs boundary correction by expanding incomplete mentions to their full span in context while filtering spurious detections that do not correspond to true PHI entities.

The Candidate Expansion loop complements this process by revisiting the entire document to identify PHI entities that were missed during the initial extraction stage. By scanning the note with awareness of the current candidate set, the system proposes additional entities that can be incorporated into the final predictions, improving recall under distribution shift and for PHI categories that are not well represented in the training data.

Together, these stages form a targeted post-processing pipeline that preserves the efficiency of NER as a high-recall candidate generator while incorporating contextual reasoning where it is most needed. This design avoids many of the deployment and interpretability challenges associated with

fully generative approaches while enabling a more controllable refinement process for clinical de-identification.

In addition to improving extraction quality, the framework is designed to support *inspection, auditability, and iterative refinement*. The system records structured information about prediction errors, including entity types, error categories, and contextual evidence, forming an error knowledge base that supports systematic analysis. The framework also generates artifacts for human review, such as fine-grained error attribution and audit-ready outputs, facilitating expert-in-the-loop validation in clinical settings.

The main technical contributions of this project are summarized as follows:

- This project develops a post-NER refinement framework with Verification and Candidate Expansion stages for improving robustness in clinical note de-identification.
- This project investigates common failure modes in NER-based de-identification systems across datasets with varying characteristics, including false positives, false negatives, and fragmented entity spans under distribution shift.
- This project designs a transparency-oriented error analysis and audit pipeline that supports structured inspection, review, and iterative refinement of de-identification results.

II. Related Work

Clinical text de-identification has evolved from early rule-based methods [1], [2] through shared benchmark tasks such as the 2014 i2b2/UTHealth de-identification challenge [3], to neural named entity recognition (NER) models [4], [5] that achieve strong performance through fine-tuning on clinical corpora. While these approaches perform well under in-distribution settings, many systems primarily expose extracted spans as their final output, providing limited support for downstream error inspection, auditing, or iterative refinement.

More recently, large language models (LLMs) have shown promise for zero-shot or instruction-based PHI removal. Approaches built on proprietary APIs such as GPT-4 [6] demonstrate strong extraction capability but require transmitting sensitive clinical text to external services, which may conflict with institutional privacy policies. Multi-stage frameworks such as RedactOR [14] improve de-identification performance through iterative refinement and verification strategies. In parallel, local and privacy-preserving LLM-based approaches [7] aim to keep inference fully on-premise while maintaining competitive performance.

Recent studies have also highlighted the importance of analyzing de-identification errors beyond aggregate metrics alone. A recent survey [15] reports that 70–80% of false positives (over-redactions) can result in clinically significant information loss, motivating the need for finer-grained error attribution and more transparent evaluation pipelines.

Building upon these directions, this project combines a locally deployed NER model with LLM-based Verification and Candidate Expansion stages using Qwen3.5-27B [9]. In addition to improving extraction robustness, the framework incorporates a transparency-oriented error analysis pipeline that associates prediction errors with specific failure modes and pipeline stages, supporting inspection and auditability in clinical de-identification workflows.

III. Personal Contributions

This report is based on a broader collaborative research project on clinical note de-identification. My primary contributions focused on system architecture, refinement pipeline design, dataset preparation, and evaluation infrastructure used throughout the project.

First, I performed preprocessing and cleaning of the i2b2 2014 benchmark dataset in order to support consistent evaluation and downstream pipeline analysis. I also constructed the synthetic dataset described in Section IV by generating clinical-note-style documents with diverse PHI patterns designed to simulate distribution shift and unseen entity categories.

Second, I designed and implemented the core post-NER refinement workflow described in Sections V-C and V-D, including both the Verification Loop and Candidate Expansion Loop. These components were developed to address key failure modes observed in NER-based de-identification systems, including fragmented entity spans, false positives, and missed PHI entities.

Third, I designed the structured Error Knowledge Base described in Section V-E, including the schema used for storing contextualized error records and retrieval-ready error metadata within Milvus. This infrastructure was developed to support downstream error analysis, retrieval-based refinement strategies, and future retrieval-augmented workflows.

In addition, I contributed to the Interactive Analytical Dashboard and evaluation infrastructure described in Sections VI-B and VI-C, including support for configurable NER backbones, refinement-stage orchestration, comparative benchmarking, and PHI-level visualization.

I also performed the primary benchmarking and analysis on the i2b2 dataset, including the performance evaluation and error-category analysis presented in Sections VII-B and VII-C.

Together, these contributions focused on building a more transparent, modular, and inspectable de-identification pipeline that supports both practical deployment and systematic error analysis.

IV. Datasets

The framework is evaluated on three datasets representing complementary evaluation settings: a standardized benchmark dataset, real-world clinical documentation, and a synthetic dataset designed to simulate distribution shift and diverse PHI patterns.

A. i2b2 2014

The i2b2 2014 de-identification dataset was introduced as part of the 2014 i2b2/UTHealth shared task on clinical text de-identification [3]. The dataset consists of longitudinal clinical narratives annotated with protected health information (PHI) entities under a detailed annotation schema. In this project, the

official test set containing 514 clinical notes is used to evaluate system performance under a standardized benchmark setting.

B. MIMIC-IV Radiology Subset

To evaluate performance on real-world clinical text, this project uses a subset of 691 radiology reports derived from the MIMIC-IV electronic health record dataset [8]. Because MIMIC-IV records are already de-identified, synthetic PHI values are injected into the notes using controlled templates to reconstruct realistic de-identification scenarios while preserving the original clinical context.

C. Synthetic Dataset

To further study robustness under varying note structures and PHI distributions, this project constructs a fully synthetic dataset generated through prompting with a large language model (Gemini 3.1 Pro). The generated notes are designed to resemble realistic clinical documentation while varying entity types, surface forms, and document structures in order to simulate distribution shift and unseen PHI patterns.

The resulting dataset contains 500 synthetic clinical notes with PHI annotations generated alongside the text to ensure consistent span labeling during evaluation.

V. Method

A. System Overview

Figure 1 illustrates the overall architecture of the proposed de-identification framework. The system extends a conventional NER-based de-identification pipeline with two post-NER refinement stages designed to improve robustness and error recovery. The framework consists of three main stages.

First, a pre-trained NER model is applied to the input clinical note to generate an initial set of PHI candidate spans. These predictions may contain both correct detections and extraction errors, including incomplete spans and spurious entities.

Second, a Verification loop examines each predicted entity within the context of the full clinical note and determines whether the span should be retained, removed, or adjusted. This stage primarily targets false positives and fragmented entity spans produced by the initial NER extraction process.

Third, a Candidate Expansion loop revisits the document to identify potential PHI entities that may have been missed during the initial extraction stage. Newly discovered candidates are incorporated into the final prediction set to improve recall under distribution shift and across diverse PHI categories.

Together, these stages form a targeted post-processing pipeline that improves extraction quality while remaining compatible with existing NER-based de-identification systems.

B. Initial NER Extraction

The de-identification pipeline begins with an NER-based extraction stage that produces an initial set of PHI candidate spans from each clinical note. Given an input document, the NER component identifies candidate spans together with their predicted entity types. The pipeline is designed with a modular extraction interface that allows different NER backbones to be integrated into the refinement workflow.

In the current evaluation setup, this project uses a RoBERTa-based de-identification model (`obi/deid_roberta_i2b2`) from the Hugging Face model hub [5] as the primary base extractor. However, the overall architecture is not tied to a specific NER implementation and is designed to support custom extraction models and alternative sequence-labeling backbones.

To obtain more reliable span-level outputs, the pipeline applies an aggregation procedure to raw token-level predictions. For models trained with BILOU-style tagging, complete entity spans are reconstructed from token labels through a standardized post-processing stage. This design helps preserve entity boundaries, reduces fragmentation, and normalizes outputs across different extraction models.

Although this stage provides an efficient first-pass PHI detector, the extracted entities are not treated as final predictions. Instead, the resulting spans are passed into the downstream Verification and Candidate Expansion stages, where additional contextual refinement and error recovery are performed.

C. Verification Loop

The Verification loop reviews the initial NER predictions and determines whether each extracted span should be preserved, removed, or adjusted. In this implementation, the stage is performed using a large language model that operates over the full clinical note together with the current set of candidate spans. Rather than re-extracting entities from scratch, the Verification loop focuses on validating and correcting candidate spans produced by the initial NER model.

A key function of this stage is boundary repair. In practice, NER models frequently return truncated or incomplete spans, particularly for names, dates, and identifier-like strings. To address this issue, the verification module examines the original clinical note together with the current detections and attempts to expand incomplete mentions to their full span in context. This step is especially important under exact-match evaluation, where partially correct spans may still be counted as errors.

In addition to boundary correction, the Verification loop also filters false positives. Given the full note and the current PHI candidate list, the verification stage identifies extracted spans that do not correspond to true PHI mentions, such as clinical terms or non-sensitive document artifacts. These spans are removed from the candidate set. After applying both boundary repair and false-positive filtering, the refined predictions are passed through a final verification step to ensure that the updated PHI set remains internally consistent.

This design enables the Verification loop to address two common NER failure modes: fragmented entity spans and incorrect extractions. By framing this stage as targeted post-validation rather than full re-extraction, the framework preserves the efficiency of NER while adding contextual reasoning where it is most beneficial.

D. Candidate Expansion Loop

While the Verification loop primarily operates on existing NER predictions, the Candidate Expansion loop is designed

to recover PHI mentions that were missed entirely during the initial extraction stage. This module revisits the full clinical note and searches for additional candidate entities that are absent from the current prediction set.

In this implementation, the expansion process is motivated by the observation that many false negatives occur because the NER model fails to propose a candidate span at all. The expansion stage therefore scans the document with access to the current extracted entities and searches for additional PHI mentions that should be incorporated into the final prediction set.

To improve efficiency, the clinical note is processed in text chunks, and newly proposed candidates from different chunks are merged and deduplicated before being added to the final output. After each round of expansion, a verification pass checks whether additional PHI entities remain undetected.

The Candidate Expansion loop directly targets false negatives and complements the Verification loop. Together, the two refinement stages transform the original NER output into a more complete and robust PHI prediction set.

E. Error Knowledge Base

To support systematic error tracking within the de-identification pipeline, the framework maintains a structured error knowledge base that records false-positive and false-negative cases produced during processing.

For each predicted entity, the system logs the PHI type, span value, surrounding context, and the corresponding verification outcome. Errors identified during the Verification and Candidate Expansion stages are categorized and stored as structured records for downstream analysis.

This design allows error information to be collected consistently across different pipeline configurations and datasets, enabling comparative analysis and iterative evaluation. The resulting error records also provide the foundation for additional analytical components, including error attribution, visualization, and retrieval-based correction workflows.

VI. Transparency and Auditability

To support trustworthy clinical deployment, the framework incorporates transparency-oriented mechanisms that make pipeline errors more inspectable and auditable. Rather than focusing solely on aggregate performance metrics, the system is designed to provide actionable diagnostic information that supports analysis, debugging, and iterative refinement of de-identification results.

A. Error Reason Attribution

Instead of reporting only aggregate false-positive and false-negative counts, the evaluation pipeline assigns a reason code to each error instance: `miss` (entity overlooked), `type` (correct span but incorrect category), `value` (boundary mismatch), and `value_type` (both incorrect boundary and category).

These error categories are mapped to different stages of the pipeline. For example, `miss` errors are often associated with the initial NER extraction stage or missed recovery during Candidate Expansion, while `value` errors frequently correspond to fragmented or incomplete entity boundaries. This structured error representation enables more targeted analysis and improvement compared to aggregate metric tuning alone.

B. Interactive Analytical Dashboard

To support downstream experimentation and analysis, the project includes an interactive analytical dashboard for configuring, benchmarking, and inspecting different de-identification pipeline variants. The dashboard is designed as a lightweight experimentation interface that allows users to evaluate multiple pipeline configurations under a unified workflow.

Users can select different NER backbones, including custom extraction models, and configure whether the Verification loop, Candidate Expansion loop, or both refinement stages are enabled during evaluation. This modular configuration design allows comparative analysis across multiple refinement strategies and model combinations without modifying the underlying pipeline implementation.

The system supports batch benchmarking across different configurations and datasets, enabling users to compare pipeline behavior under varying extraction and refinement settings. Evaluation results are visualized through interactive charts and metric dashboards, including precision, recall, F1, false-positive rates, and entity-level performance breakdowns.

To support note-level inspection, the interface also provides inline PHI highlighting and color-coded annotations for extracted entities and error categories. This visualization layer allows users to inspect how different refinement stages affect prediction behavior and error recovery within individual clinical notes.

C. Compliance-Oriented Offline Auditability

To support offline review in clinical and compliance settings, the evaluation pipeline generates structured analysis reports containing aggregated metrics, configuration summaries, note-level performance breakdowns, and contextual error summaries. These reports are designed to support expert review, manual inspection, and iterative refinement of de-identification results without requiring direct access to the running system.

For each evaluation run, the system exports structured spreadsheet reports containing multiple analysis views, including overall benchmark summaries, per-note prediction results, PHI-type breakdowns, and categorized false-positive and false-negative error analyses. The generated reports also record the evaluated pipeline configuration, including the selected NER backbone and whether the Verification and Candidate Expansion stages were enabled during inference.

To support detailed inspection, the exported artifacts include contextual information for each prediction error, such as the predicted entity span, ground-truth label, error category, and surrounding clinical text context. This design allows reviewers to trace aggregate metrics back to individual extraction failures and inspect how different refinement stages affect prediction behavior at the note level.

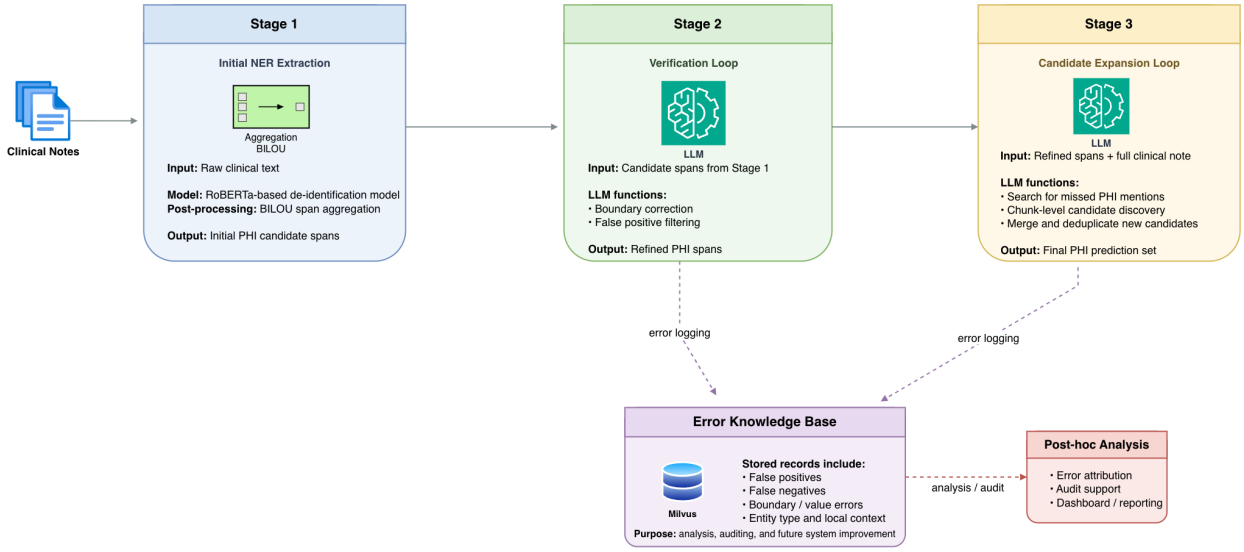

Fig. 1. Architecture of the proposed de-identification framework. A base NER model first extracts candidate PHI entities from clinical notes. Two refinement loops are then applied: a verification loop that validates and fixes extracted entities and a candidate expansion loop that discovers missed PHI mentions. Errors identified during the process are recorded in an error knowledge base for analysis.

In addition, error records can be stored in a vector database for retrieval-based analysis and future correction workflows. By embedding historical error cases together with their contextual information, the system can support future retrieval-augmented refinement strategies for recurring failure patterns and difficult PHI categories under distribution shift.

VII. Experimental Evaluation

A. Experimental Setup

The de-identification pipeline uses a RoBERTa-based clinical de-identification model (`obi/deid_roberta_i2b2`) [5] as the NER backbone together with a sliding-window BILOU decoder (`max_length`=512, `stride`=128).

The evaluation compares three configurations: **OBI-RoBERTa** [5] as the baseline model; **+ Verification**, which applies only the Verification Loop to reduce false positives and repair fragmented spans; and **+ Verif. + Cand. Exp.**, which applies both the Verification and Candidate Expansion stages to recover missed PHI entities.

All LLM inference uses Qwen3.5-27B [9] served locally through vLLM [10] on an institutional HPC cluster. Running inference locally allows the pipeline to process PHI-sensitive clinical text without transmitting data to external cloud services.

Evaluation is performed at both the token and entity levels. *Token-level* metrics (Table I) follow the i2b2 [3] evaluation protocol, which tokenizes PHI spans into alphanumeric tokens and awards partial credit for overlapping detections through set intersection. These metrics are reported in both binary (any PHI) and category-aware (entity type must also match) settings.

Entity-level exact-match metrics (Table II) require both the extracted span value and PHI type to exactly match a ground-truth entity. Under this setting, type mismatches count as both false positives and false negatives, while partial span overlaps receive no credit.

B. Overall Performance

In clinical de-identification, false negatives represent a particularly important safety concern because missed PHI entities allow sensitive information to remain in released clinical text. For this reason, the evaluation emphasizes both entity-level F1 and false negative rate (FNR).

On the **synthetic dataset**, the full refinement pipeline produces the largest improvements. Entity-level F1 increases from 61.52% to 87.78%, while false positives are reduced by 77.6% (1780 $\rightarrow$ 399) and false negatives by 61.0% (911 $\rightarrow$ 355). The synthetic dataset contains 14 PHI types, including EMAIL, SSN, and ACCESSION, that are not present in the i2b2 training corpus. The Candidate Expansion stage is able to recover many of these entities without additional supervised training, demonstrating improved robustness under distribution shift.

On the **i2b2 benchmark**, where the baseline NER model is already well calibrated (entity F1 = 84.00%), the Verification loop improves precision (87.59%$\rightarrow$89.92%), while the full refinement pipeline reduces entity-level F1 to 80.83% because of increased false positives. However, the full pipeline achieves the lowest FNR (18.42%), suggesting that the refinement stages favor recall-oriented recovery in safety-sensitive settings.

On the **MIMIC-IV radiology subset**, the full pipeline improves entity-level F1 from 84.58% to 86.50% while reducing FNR from 16.48% to 13.90%. Applying the Verification loop alone also reduces false positives by 19.4% (211$\rightarrow$170), indicating that boundary repair and false-positive filtering remain effective on real-world clinical text.

C. Error Analysis by PHI Type

Table III presents a breakdown of errors by PHI category, highlighting two major failure modes targeted by the refinement stages.

TABLE I
TOKEN-LEVEL DE-IDENTIFICATION PERFORMANCE ACROSS DATASETS

Dataset	Configuration	Token (Binary)			Token (Category)		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
i2b2 2014	OBI-RoBERTa [5]	96.64	91.74	**94.13**	92.63	**87.93**	**90.22**
	+ Verification	**97.56**	89.48	93.35	**94.00**	86.21	89.94
	+ Verif. + Cand. Exp.	90.86	**92.45**	91.65	85.84	87.35	86.59
MIMIC-IV Radiology Subset	OBI-RoBERTa	97.05	86.44	91.43	96.85	86.27	91.25
	+ Verification	**99.14**	86.33	**92.29**	**98.98**	86.20	**92.15**
	+ Verif. + Cand. Exp.	96.70	**88.08**	92.19	96.55	**87.94**	92.04
Synthetic Dataset	OBI-RoBERTa	85.41	87.78	86.58	72.49	74.50	73.48
	+ Verification	96.17	76.55	85.24	86.20	68.62	76.41
	+ Verif. + Cand. Exp.	**98.16**	**94.46**	**96.27**	**92.45**	**88.96**	**90.67**

TABLE II
ENTITY-LEVEL (EXACT MATCH) DE-IDENTIFICATION PERFORMANCE ACROSS DATASETS

Dataset	Configuration	P (%)	R (%)	F1 (%)	FP	FN	FNR (%)
i2b2 2014	OBI-RoBERTa [5]	87.59	80.70	**84.00**	1311	2212	19.30
	+ Verification	**89.92**	78.30	83.71	**1006**	2487	21.70
	+ Verif. + Cand. Exp.	80.09	**81.58**	80.83	2324	**2111**	**18.42**
MIMIC-IV Radiology Subset	OBI-RoBERTa	85.68	83.52	84.58	211	249	16.48
	+ Verification	**88.12**	83.45	85.72	**170**	250	16.55
	+ Verif. + Cand. Exp.	86.91	**86.10**	**86.50**	196	**210**	**13.90**
Synthetic Dataset	OBI-RoBERTa	54.72	70.25	61.52	1780	911	29.75
	+ Verification	75.09	67.15	70.90	682	1006	32.85
	+ Verif. + Cand. Exp.	**87.15**	**88.41**	**87.78**	**399**	**355**	**11.59**

TABLE III
ERROR DISTRIBUTION BY PHI TYPE ON THE SYNTHETIC DATASET

PHI Type	OBI-RoBERTa [5]				+ Verif. + Cand. Exp.				FP+FN Reduction (%)
	FP	(%)	FN	(%)	FP	(%)	FN	(%)	
MRN	991	(55.7)	242	(26.6)	107	(26.8)	53	(14.9)	87.0
DATE	392	(22.0)	8	(0.9)	80	(20.1)	3	(0.8)	79.3
PHONE	137	(7.7)	–	–	33	(8.3)	–	–	75.9
NAME	98	(5.5)	69	(7.6)	71	(17.8)	69	(19.4)	16.2
HOSPITAL	60	(3.4)	–	–	36	(9.0)	–	–	40.0
EMAIL	7	(0.4)	52	(5.7)	1	(0.3)	3	(0.8)	93.2
ADDRESS	3	(0.2)	–	–	54	(13.5)	–	–	↑
ACCESSION	–	–	294	(32.3)	–	–	62	(17.5)	78.9
ORGANIZATION	–	–	60	(6.6)	–	–	60	(16.9)	0.0
USERNAME	–	–	63	(6.9)	–	–	28	(7.9)	55.6
LOCATION	–	–	48	(5.3)	–	–	37	(10.4)	22.9
BIOMETRIC	–	–	35	(3.8)	–	–	35	(9.9)	0.0
SSN	–	–	26	(2.9)	–	–	1	(0.3)	96.2
OTHER	92	(5.2)	14	(1.5)	17	(4.3)	4	(1.1)	80.2
Total	**1780**	**(100)**	**911**	**(100)**	**399**	**(100)**	**355**	**(100)**	**72.0**

Complete entity overlook (`miss`) produces false negatives without corresponding false positives. This failure mode appears in PHI categories such as ACCESSION, USERNAME, ORGANIZATION, LOCATION, BIOMETRIC, and SSN, where

the baseline NER model frequently fails to propose candidate spans.

In contrast, *boundary or type mismatch* (`value`/`type`) produces both false positives and false negatives for the same entity type. These errors appear most prominently in MRN (991 FP / 242 FN) and DATE (392 FP / 8 FN), where fragmented spans and incorrect boundaries are common.

The Candidate Expansion stage primarily targets `miss` errors, reducing total errors by 96.2% on SSN and 93.2% on EMAIL. The Verification loop primarily targets `value`/`type` errors, reducing MRN false positives from 991 to 107 and DATE false positives from 392 to 80 through boundary repair and false-positive filtering.

D. Discussion

Across all three datasets, the Verification loop improves precision while the Candidate Expansion loop recovers missed entities. On the synthetic dataset, FNR decreases from 29.75% to 11.59% (over 2.5$\times$ reduction), while on the i2b2 benchmark, the Verification loop alone removes 23.3% of false positives.

Although aggregate metrics and category-level breakdowns (Table III) quantify the effectiveness of the pipeline, the transparency mechanisms described in Section VI provide additional context for understanding why these improvements occur. Linking quantitative gains back to specific failure modes, such as repairing fragmented boundaries or recovering overlooked identifiers, makes it possible to analyze how different refinement stages affect overall system behavior.

The evaluation results suggest that the framework is most effective as an interpretable and auditable refinement layer for existing NER systems rather than as a complete replacement for domain-specific fine-tuning. In clinical de-identification settings, this refinement-oriented design provides a practical balance between extraction quality, controllability, and deployment transparency.

VIII. Conclusion

This project presented a post-NER refinement framework for clinical text de-identification that augments conventional NER pipelines with two complementary stages: Verification and Candidate Expansion. Rather than replacing the underlying NER model, the framework treats NER extraction as a high-recall candidate generation stage and applies targeted refinement to address key failure modes, including false positives, false negatives, and fragmented entity spans.

The evaluation highlights several important observations. First, post-NER refinement improves robustness under distribution shift, where conventional NER models often struggle with variability in PHI patterns and document structure. Second, the Verification and Candidate Expansion stages address complementary error modes: Verification primarily improves precision through boundary repair and false-positive filtering, while Candidate Expansion improves recall by recovering overlooked PHI entities.

The results also demonstrate the importance of transparency and auditability in clinical de-identification workflows. By structuring error information and exposing it through analysis and reporting mechanisms, the framework supports systematic inspection, debugging, and iterative refinement of de-identification results. In safety-sensitive settings such as healthcare, this refinement-oriented design provides a practical balance between extraction quality, interpretability, and deployment control.

Overall, the project demonstrates that combining post-NER refinement with transparency-oriented analysis mechanisms provides a practical foundation for building more robust and inspectable clinical de-identification systems.

IX. Future Work

Several directions remain for future exploration and system extension.

One promising direction is to integrate the structured error knowledge base into both inference and training workflows. Retrieved historical error cases could guide refinement decisions during inference while also supporting targeted fine-tuning for recurring failure patterns. In particular, difficult PHI categories and frequently fragmented entity patterns could be used to construct harder supervision examples for improving the initial NER extraction stage.

Another important direction is the development of retrieval-augmented refinement strategies. Instead of performing verification solely on the current clinical note, future versions of the system could retrieve similar historical error cases from the vector database and use them as contextual references during refinement. Such a retrieval-based strategy may improve robustness for rare PHI categories, unseen entity formats, and highly variable clinical note structures under distribution shift.

The structured error knowledge base may also support the construction and evaluation of more challenging synthetic datasets. By analyzing which PHI categories, boundary patterns, or note structures consistently produce errors, future work could generate synthetic benchmark datasets that explicitly target known failure modes. This would allow more systematic evaluation of de-identification robustness beyond conventional benchmark settings.

Another practical extension is to integrate the framework into real-time clinical document processing or streaming systems [16]. Supporting continuous, low-latency de-identification for incoming clinical notes would enable more realistic deployment scenarios for privacy-preserving clinical data processing pipelines.

Finally, future work could explore tighter integration between the refinement stages and the underlying NER model. Instead of treating refinement solely as a post-processing step, verified corrections and recovered entities could be fed back into iterative model training, allowing the extraction backbone to gradually internalize common failure patterns and reduce errors during the initial extraction stage itself.

Acknowledgment

This report is based on collaborative research conducted with Ruoqian Zhang, whose contributions to the broader collaborative development of the project are gratefully acknowledged.

This work is supported by a Brown Seed Fund.

REFERENCES

[1] L. Sweeney, "Replacing personally-identifying information in medical records, the Scrub system," *Proc. AMIA Annual Fall Symposium*, pp. 333–337, 1996.

[2] I. Neamatullah *et al.*, "Automated de-identification of free-text medical records," *BMC Medical Informatics and Decision Making*, vol. 8, no. 32, 2008.

[3] A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1," *J. Biomedical Informatics*, vol. 58, pp. S11–S19, 2015.

[4] F. Dernoncourt *et al.*, "De-identification of patient notes with recurrent neural networks," *J. American Medical Informatics Association*, vol. 24, no. 3, pp. 596–606, 2017.

[5] P. Kailas *et al.*, "Robust de-identification of medical notes using transformer architectures," Zenodo, 2022. [Online]. Available: https://zenodo.org/records/6617957

[6] Z. Liu *et al.*, "DeID-GPT: Zero-shot medical text de-identification by GPT-4," *arXiv preprint arXiv:2303.11032*, 2023.

[7] I. C. Wiest *et al.*, "Deidentifying medical documents with local, privacy-preserving large language models: The LLM-Anonymizer," *NEJM AI*, vol. 2, no. 4, 2025.

[8] A. E. W. Johnson *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, 2023.

[9] A. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[10] W. Kwon *et al.*, "Efficient memory management for large language model serving with PagedAttention," in *Proc. ACM SIGOPS 29th Symp. Operating Systems Principles*, 2023.

[11] J. Wang *et al.*, "Milvus: A purpose-built vector data management system," in *Proc. 2021 Int. Conf. Management of Data (SIGMOD)*, pp. 2614–2627, 2021.

[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP*, 2019.

[13] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, 2020.

[14] P. Singh *et al.*, "RedactOR: An LLM-powered framework for automatic clinical data de-identification," in *Proc. ACL 2025 Industry Track*, 2025.

[15] K. Aghakasiri *et al.*, "Not what the doctor ordered: Surveying LLM-based de-identification and quantifying clinical information loss," in *Proc. EMNLP*, 2025.

[16] S. Chen *et al.*, "VectraFlow: Long-horizon semantic processing over data and event streams with LLMs," *arXiv preprint arXiv:2604.03855*, 2026.