Improving Clinical Note De-identification via Post-NER Verification and Candidate Expansion

Ruoqian Zhang
Brown University
ruoqian_zhang@brown.edu

***Abstract*—De-identification of clinical notes is critical for protecting patient privacy, yet existing approaches often struggle under real-world variation and provide limited support for auditing and error analysis. By examining the outputs of NER-based systems, we observe three recurring failure modes: false positives, false negatives, and fragmented entity spans. The latter can simultaneously introduce both error types under exact-match evaluation.**

We present a transparent, locally deployable de-identification framework that augments NER-based extraction with two refinement stages: a Verification loop to correct candidate entities and a Candidate Expansion loop to recover missed protected health information (PHI). Beyond improving extraction quality, the system generates structured artifacts for human review, including fine-grained error attribution, audit-ready spreadsheet exports, and an interactive analytical dashboard for cross-configuration comparison, enabling systematic inspection and iterative refinement of de-identification results.

We evaluate the framework on the i2b2 2014 benchmark, a MIMIC-IV radiology subset, and a synthetic dataset simulating distribution shift. Results show substantial robustness gains under variation, increasing entity-level F1 from 61.52% to 87.78% and reducing the false-negative rate from 29.75% to 11.59% on the synthetic dataset, while maintaining competitive performance on benchmark data. These findings highlight the value of combining post-NER refinement with transparency-oriented evaluation infrastructure for reliable clinical de-identification.

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 settings, such variability is common, raising concerns about the robustness and reliability of these systems. Moreover, most de-identification pipelines operate as black boxes, providing limited support for understanding, auditing, or correcting errors, capabilities that are essential in clinical and compliance workflows.

To better understand these limitations, we analyze 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. Our analysis shows that the same model exhibits 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. Notably, fragmented spans are particularly problematic under strict span-level evaluation, as they can simultaneously introduce both false positives and false negatives.

To address these challenges, we introduce 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. The framework treats the output of the NER model as an initial set of candidate spans and applies targeted refinement rather than re-extracting entities from scratch.

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

The candidate expansion stage complements this process by revisiting the entire document to identify PHI entities that were missed during initial extraction. By scanning the note with awareness of the current candidate set, the system proposes additional entities that are then incorporated into the final predictions, improving recall, especially under distribution shift or when encountering PHI types 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 the cost and opacity of fully generative approaches, enabling a controllable and interpretable refinement process suitable for practical deployment in clinical settings.

Beyond improving extraction quality, our 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 enables systematic analysis. In addition, the framework produces artifacts for human review, including fine-grained error attribution, audit-ready spreadsheet exports, and an interactive analytical dashboard for comparing pipeline configurations and inspecting note-level errors, facilitating expert-in-the-loop validation in clinical settings.

In summary, this work makes the following contributions:

- We propose a post-NER refinement framework with Verification and Candidate Expansion stages that addresses common de-identification errors while maintaining practical deployment constraints.
- We develop an evaluation methodology with structured error attribution that classifies each prediction error into one of four categories (`miss`, `type`, `value`, `value_type`) and maps it to a specific pipeline stage, enabling systematic analysis of failure modes across three datasets with varying characteristics.
- We build a transparency-oriented audit infrastructure, comprising fine-grained error reasoning, audit-ready spreadsheet reports, and an interactive analytical dashboard for cross-configuration comparison that supports inspection, debugging, and iterative refinement of clinical de-identification pipelines.

II. RELATED WORK

Clinical text de-identification has evolved from early rule-based methods [1], [2] through shared benchmarks such as the 2014 i2b2/UTHealth shared task [3], to neural NER models [4], [5] that achieve high accuracy through fine-tuning on clinical corpora. While effective in distribution, these systems typically expose detected spans as their primary output, leaving downstream error diagnosis to ad-hoc tooling.

More recently, LLMs have shown promise for zero-shot PHI removal. Approaches built on proprietary APIs such as GPT-4 [6] achieve strong accuracy but require transmitting sensitive text to external servers, which can be incompatible with institutional policy. Multi-stage frameworks such as RedactOR [14] report strong i2b2 results through iterative detection. Local, privacy-preserving LLMs [7] keep data on-premise and report competitive aggregate metrics. Complementing these efforts, a recent survey [15] highlights that 70–80% of false positives (over-redactions) cause clinically significant information loss, motivating finer-grained error attribution alongside aggregate scores.

Our work is complementary to this line of research: we combine a locally deployed NER model with an LLM-based verification and expansion stage (using Qwen3.5-27B [9]) and add a transparency layer that attributes each error to a specific failure mode and pipeline stage, supporting auditability in clinical deployment.

III. PERSONAL CONTRIBUTIONS

My personal focus throughout this project has been on the evaluation methodology and the transparency-oriented infrastructure that surround the refinement pipeline—the components that turn raw model output into reviewable evidence. The framework was developed jointly with Zejun Zhou, whose focus was on system design and the LLM-based refinement loops; my own work targets the half of the paper's argument that transparency-oriented evaluation infrastructure is what makes such a pipeline auditable in practice.

Three threads of this work directly motivate the discussion that follows. First, I designed and implemented the four-category reason-code attribution scheme (`miss`, `type`, `value`, `value_type`) together with the strict type-matching policy described in Section VI-A; this scheme converts aggregate error counts into stage-attributable diagnoses and underlies the cross-dataset error analysis in Section VII-C. Second, I built the audit-ready Excel export pipeline with auto-generated reasoning fields, described in Section VI-C, which supports the compliance-oriented review workflow foregrounded in the discussion in Section VII-E. Third, I designed and implemented the interactive analytical dashboard described in Section VI-B, including its configuration-comparison panels, drill-down views, and note-level error highlighting.

Beyond these infrastructure contributions, I also carried out the experimental evaluation itself: running the three pipeline configurations on the i2b2 2014, MIMIC-IV radiology, and synthetic datasets, and producing the per-PHI-type failure-mode analysis. Taken together, these contributions reflect a consistent emphasis: that a de-identification pipeline should be judged not only by its aggregate metrics, but by how legible and reviewable its errors are.

IV. DATASETS

We evaluate the proposed framework on three datasets representing complementary evaluation settings: a standard 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 our experiments, we use the official test set containing 514 clinical notes to evaluate model performance under a standardized benchmark setting.

B. MIMIC-IV Radiology Subset

To evaluate performance on real-world clinical text, we use 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 model robustness under different note structures and PHI distributions, we construct a fully synthetic dataset generated through prompting with a large language model (Gemini 3.1 Pro). The generated notes are designed to resemble typical clinical documentation while varying entity types, surface forms, and document structures. The resulting dataset contains 500 notes with PHI annotations generated alongside the text to ensure consistent span labeling.

V. Method

A. System Overview

Figure 1 illustrates the overall architecture of the proposed framework. Our framework augments a conventional NER-based de-identification pipeline with two post-NER refinement stages. The framework consists of three 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.

Second, a verification loop examines each predicted entity and determines whether the span should be retained, removed, or adjusted. This step aims to reduce false positives and repair incorrect boundaries produced by the initial NER model.

Third, a candidate expansion loop revisits the entire document to identify potential PHI entities that may have been missed in the initial extraction stage. Newly discovered candidates are then incorporated into the final prediction set.

Together, these stages form a refinement process that improves extraction quality while remaining compatible with existing NER-based systems.

B. Initial NER Extraction

The proposed framework begins with an NER-based extraction stage that produces an initial set of PHI candidate spans from each clinical note. Given the input text, the NER model identifies spans together with their predicted entity types. In our experiments, we use a RoBERTa-based de-identification model (`obi/deid_roberta_i2b2`) from the Hugging Face model hub [5] as the base extractor. These predictions serve as input to the subsequent refinement stages.

To obtain more reliable span-level outputs, we apply an aggregation procedure to raw token-level predictions. For models trained with BILOU-style tagging, we reconstruct complete entity spans from token labels. This step follows standard sequence labeling post-processing and helps preserve boundaries, reducing fragmentation in the initial results.

Although this stage provides an efficient first-pass detector, its outputs are not used as final predictions. Instead, the extracted spans are treated as candidates for the subsequent verification and expansion stages.

C. Verification Loop

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

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

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

This design allows the verification loop to address two of the most common NER failure modes, which are fragmented spans and incorrect extractions. By framing this stage as post validation rather than full re-extraction, the framework preserves the efficiency of NER while adding contextual reasoning where it is most needed.

D. Candidate Expansion Loop

While the verification loop operates primarily 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 our implementation, the expansion process is guided by the observation that many false negatives arise because the NER model fails to propose a candidate span at all. The agent therefore scans the entire document with access to the current extracted entities and looks for additional PHI occurrences that should also be included.

To improve efficiency, the note is processed in text chunks, and newly proposed candidates from different chunks are merged and deduplicated before being added to the final prediction set. After each round of expansion, a verification pass checks whether any additional PHI is still missing.

The candidate expansion loop directly targets false negatives and complements the verification loop. Together, the two modules transform the original NER output into a more complete and better-calibrated PHI prediction set.

E. Error Knowledge Base

To support systematic error tracking within the pipeline, we maintain a structured error knowledge base that records false positive and false negative cases produced during processing.

For each predicted entity, the system logs its 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.

This design allows error information to be consistently collected across different pipeline configurations and datasets, enabling downstream analysis and comparison. The stored error records serve as the foundation for further analytical components, including error attribution, visualization, and retrieval-based correction.

VI. Transparency and Auditability

To support trustworthy clinical deployment, our system incorporates mechanisms that make pipeline errors transparent and auditable, redirecting attention from aggregate performance metrics to actionable diagnostic insights.

A. Error Reason Attribution

Rather than reporting bare counts, the evaluation module attaches a reason code to each error: `miss` (entity overlooked), `type` (correct span, wrong category), `value` (boundary mismatch), or `value_type` (both wrong). These codes map to pipeline stages: `miss` points to NER or Candidate Expansion, `type` to category assignment, `value` to boundary detection, and `value_type` to compound failures—enabling targeted improvements rather than blind tuning.

In practice, each prediction is matched against ground-truth annotations under a strict type-matching policy: when a predicted span and a ground-truth entity share boundaries, a category disagreement is recorded as a `type` error rather than treated as a partial match, preventing optimistic aggregates from masking systematic category drift. Boundary- level mismatches are recorded separately from category-level ones so that each error contributes to a single, actionable diagnosis tied to the pipeline stage most likely responsible for it. The same matching policy is applied uniformly across the three datasets, which is what makes the cross-dataset error distributions reported in Section VII directly comparable.

B. Interactive Analytical Dashboard

To operationalize this transparency, we implemented an interactive analytical dashboard for multi-dimensional error analysis. Users define pipeline configurations by selecting the NER backbone, enabling or disabling the Verification and Candidate Expansion loops, and adjusting LLM parameters such as temperature and chunk size, and evaluate them on the same benchmark dataset. Results are rendered as grouped bar charts across precision, recall, and F1 at both token and entity levels, with per-PHI-type error breakdowns surfaced as drill-down panels for direct comparison of how each refinement stage affects specific error categories.

At the note level, the dashboard renders the original clinical text with color-coded inline highlighting for each PHI category, overlaying false-positive and false-negative markers linked to their reason codes for in-context inspection. To accommodate large-scale evaluations, the system checkpoints intermediate state after each processed note, allowing benchmarks to resume after interruption and supporting partial result visualization while a run is in progress—behavior that decouples evaluation execution from result review and lets reviewers iterate on configurations without waiting for full benchmarks to complete.

C. Compliance-Oriented Offline Auditability

In clinical settings governed by HIPAA and institutional review requirements, non-technical stakeholders, including compliance officers and medical experts, must be able to audit de-identification outcomes without access to the running system. Upon completion of each evaluation, the system generates structured Excel reports with aggregated metrics, note-level performance breakdowns, and error listings with reason codes and context, enabling expert-in-the-loop validation and targeted feedback.

To support per-error review, every false-positive and false-negative entry is augmented with an automatically generated reasoning field that links the predicted span (or missed annotation) to the assigned reason code, the matching ground-truth entity, and the immediate clinical context window in which the error occurred. This pairs each aggregate-level metric with a unit of evidence that reviewers can inspect, dispute, or accept, closing the loop between numerical evaluation and qualitative auditing.

In addition, each error record is stored as a 384-d embedding (`all-MiniLM-L6-v2` [12]) in Milvus [11], forming a queryable error corpus that supports retrieval- augmented correction [13] in future work.

VII. Experimental Evaluation

A. Experimental Setup

The NER backbone is a RoBERTa-based de-identification model (`obi/deid_roberta_i2b2`) [5] with a sliding-window BILOU decoder (`max_length`=512, `stride`=128). We compare three configurations: **OBI-RoBERTa** [5] (baseline); **+ Verification**, applying only the Verification Loop to filter false positives; and **+ Verif. + Cand. Exp.**, the full pipeline that first executes the Verification Loop, then applies Candidate Expansion to recover missed entities. All LLM inference uses Qwen3.5-27B [9] served locally via vLLM [10] on an institutional HPC cluster, ensuring PHI never leaves the institutional network: a key advantage over cloud-based approaches such as DeID-GPT [6].

We evaluate at two levels. *Token-level* metrics (Table I) follow the i2b2 [3] protocol, which tokenizes each PHI span into alphanumeric tokens and awards partial credit for overlapping detections via set intersection, reported in both binary (any PHI) and category-aware (type must also match) variants. *Entity-level* exact-match metrics (Table II) require both the extracted value and PHI type to match a ground-truth

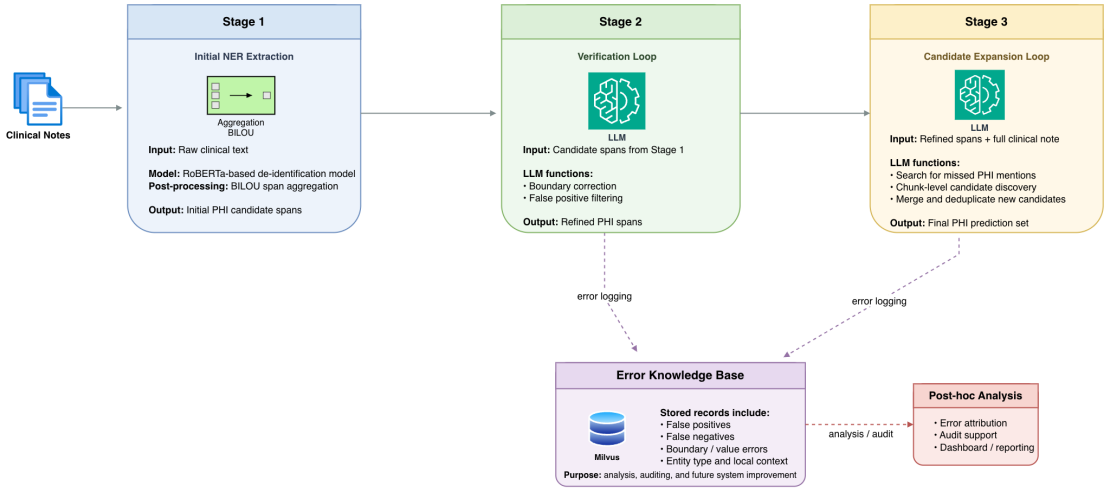

Fig. 1. Architecture of the proposed de-identification framework. A base NER model first extracts candidate PHI entities from clinical notes. Two refinement loops are then applied: a verification loop that validates and fixes extracted entities and a candidate expansion loop that discovers missed PHI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

entity; a type mismatch counts as both a false positive and a false negative, and partial span overlaps receive no credit.

B. Overall Performance

In clinical de-identification, false negatives pose a greater risk than false positives: a missed PHI entity means sensitive information remains in released data, constituting a direct privacy violation. We therefore treat **FNR** as the primary safety metric alongside F1.

On the **synthetic dataset**, the full pipeline yields the largest gains: entity F1 rises from 61.52% to 87.78%, with FP reduced by 77.6% (1780 $\rightarrow$ 399) and FN by 61.0% (911 $\rightarrow$ 355). The synthetic dataset contains 14 PHI types including EMAIL, SSN, and ACCESSION, which are absent from the i2b2 training corpus. The Candidate Expansion loop detects these entities zero-shot, providing extensibility without additional training cost.

On the **i2b2 benchmark**, where NER is already well-calibrated (entity F1 = 84.00%), the Verification loop improves precision (87.59%$\rightarrow$89.92%) but the full pipeline decreases entity F1 to 80.83% due to increased false positives. However, it achieves the lowest FNR (18.42%), indicating the agent may over-detect on well-studied benchmarks but remains valuable for safety.

On the **MIMIC-IV radiology subset**, the full pipeline improves entity F1 from 84.58% to 86.50% and reduces FNR from 16.48% to 13.90%, with verification alone cutting false positives by 19.4% (211$\rightarrow$170).

C. Error Analysis by PHI Type

Table III breaks down errors by PHI type, revealing two distinct failure modes addressed by the two loops.

Complete entity overlook (`miss`) produces FN without FP, affecting ACCESSION, USERNAME, ORGANIZATION, LOCATION, BIOMETRIC, and SSN (all zero baseline FP).

Boundary or type mismatch (`value`/`type`) produces both FP and FN on the same type, dominating MRN (991 FP / 242 FN) and DATE (392 FP / 8 FN).

The Candidate Expansion loop targets `miss`, reducing total errors by 96.2% on SSN and 93.2% on EMAIL. The Verification Loop targets `value`/`type`, cutting MRN FP from 991 to 107 and DATE FP from 392 to 80.

D. Case Examples

Case 1: Recovering `miss` Errors (i2b2, Item 50). Physician names "XIAN" and "QUINLIVAN" in clinical shorthand ("NEG BY XIAN") are missed by NER. The agent recovers both through contextual reasoning.

Case 2: Correcting `value` Errors (synthetic dataset, Item 25). NER fragments MRN "384-92-7476", accession "2026.02.10.485", and date "2026-06-15". The Verification Loop reconstructs the complete spans from these fragments, reducing 3 FN + 4 FP to 0 FN + 0 FP.

E. Discussion

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

While aggregate metrics and category-level breakdowns (Table III) quantify the pipeline's effectiveness, interpreting these results through our transparency mechanisms (Section VI) provides the crucial context of *why* these improvements occur. By tracing quantitative gains back to specific qualitative failure modes, such as fixing boundary mismatches or recovering completely overlooked identifiers, we establish that this pipeline is best deployed not as a black-box replacement for domain-specific fine-tuning, but as an interpretable, auditable safety net for clinical data release.

TABLE I
TOKEN-LEVEL DE-IDENTIFICATION PERFORMANCE ACROSS DATASETS

		Token (Binary)			Token (Category)		
Dataset	**Configuration**	**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

	OBI-RoBERTa [5]				+ Verif. + Cand. Exp.				
PHI Type	**FP**	**(%)**	**FN**	**(%)**	**FP**	**(%)**	**FN**	**(%)**	**FP+FN Reduction (%)**
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**

VIII. CONCLUSIONS AND FUTURE WORK

We 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, our approach treats it as a high-recall candidate generator and applies targeted refinement to address key failure modes, including false positives, false negatives, and fragmented entity spans.

Our study highlights three main takeaways. First, post-NER refinement improves robustness under distribution shift, where standard NER models struggle with variability in PHI patterns and document structure. Second, verification and expansion address complementary error modes: verification improves precision through boundary correction and false-positive filtering, while expansion improves recall by recovering missed entities, yielding a controllable precision–recall trade-off. In clinical settings, false negatives are especially critical, as each missed PHI instance represents a potential HIPAA violation, while false positives primarily reduce the utility of the data.

Third, transparency and auditability are essential for real-world deployment. By structuring error information and exposing it through analysis and reporting, the framework enables systematic inspection, debugging, and iterative improvement, supporting expert review, defensible risk assessment, and compliance-oriented workflows.

These results suggest that effective de-identification systems should be designed not only for accuracy but also for transparency. In particular, lightweight extraction combined with targeted contextual refinement offers a compelling alternative to fully generative approaches, balancing performance, cost, and control.

Looking ahead, a promising direction is to integrate the error knowledge base into both inference and training. Retrieved error cases can guide refinement decisions at inference time and support targeted fine-tuning, enabling the system to internalize recurring failure patterns. Additionally, integrating the pipeline into real-time clinical document streaming frameworks [16] would allow continuous on-the-fly removal of PHI.

Overall, this work demonstrates that combining post-NER refinement with structured transparency mechanisms provides a practical foundation for building robust, interpretable, and compliance-ready clinical de-identification systems.

ACKNOWLEDGMENT

This report is based on collaborative research conducted with Zejun Zhou, whose contributions to the system design and method implementation 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.