Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 2 (2025)

Article

Intelligent Detection and Protection of Personally Identifiable Information in Clinical Text: An Advanced NLP Approach with Optimized Attention Mechanisms

Haoyang Guan [1,*]

[1] Data Science, Columbia University, NY, USA
* Correspondence: Haoyang Guan; Data Science, Columbia University, NY, USA

Abstract: The protection of Personally Identifiable Information (PII) in clinical text is a critical challenge in healthcare data management, particularly as medical institutions increasingly adopt digital health records and data-sharing initiatives. This paper presents a novel natural language processing framework that leverages optimized attention mechanisms and context-aware tokenization strategies to achieve high accuracy in detecting and protecting sensitive information within clinical documents. Our approach integrates transformer-based architectures with domain-specific enhancements, achieving a 95.3% F1-score on standard benchmarks while satisfying HIPAA Safe Harbor requirements through a combination of deep learning and rule-based processing. The proposed method introduces a hierarchical detection system that processes clinical text at multiple granularity levels, employing specialized attention heads for different PII categories. Experimental results on three large-scale clinical datasets demonstrate that our framework outperforms existing state-of-the-art methods by 8.7% in detection accuracy and reduces false positives by 59% compared to ClinicalBERT (from 12.8% to 5.2%). Furthermore, our intelligent redaction strategy preserves the semantic integrity of clinical content, enabling secure data sharing while maintaining the utility of medical information.

Keywords: clinical text de-identification; personally identifiable information detection; attention mechanism; HIPAA compliance

Received: 15 September 2025
Revised: 01 December 2025
Accepted: 16 December 2025
Published: 22 December 2025

1. Introduction

1.1. Healthcare Data Privacy Challenges

The digitalization of healthcare systems has generated unprecedented volumes of clinical text data, including electronic health records, physician notes, discharge summaries, and consultation reports. These documents contain rich medical information essential for patient care, clinical research, and healthcare analytics. The need for secure data sharing in multicenter research studies has driven the development of various de-identification and anonymization strategies. However, they also contain sensitive Personally Identifiable Information (PII) that must be rigorously protected to maintain patient privacy and comply with regulatory requirements.

Critical Limitations of Existing Methods

Current state-of-the-art approaches exhibit three main limitations:

1) Uniform attention heads that fail to specialize by PII type, such as temporal versus identity information.
2) Single-granularity processing that overlooks multi-scale patterns spanning tokens, segments, and documents.

3) Binary redaction methods, like those used in ClinicalBERT, that treat all attention heads uniformly and fail to preserve contextually important information.

Flat Processing: Existing approaches process text at a single granularity level, missing patterns that occur across multiple scales.

Binary Redaction: Traditional methods employ all-or-nothing redaction, which can destroy clinical utility by removing contextually significant information. While the Health Insurance Portability and Accountability Act (HIPAA) mandate the removal of 18 specific identifier categories, automated systems currently achieve only an 85-88% F1-score on comprehensive evaluations. Manual de-identification costs healthcare institutions approximately $0.80-$1.20 per document, totaling billions of dollars annually across the US healthcare system [1]. Previous studies have demonstrated that traditional de-identification approaches can significantly reduce the informational content and utility of clinical documents, highlighting the need for intelligent methods that balance privacy protection with data utility [2].

The de-identification of unstructured clinical data presents unique challenges compared to structured data, requiring advanced natural language processing techniques to accurately identify and protect sensitive information while preserving clinical meaning [3]. Our framework addresses these challenges through three key innovations: specialized attention heads for PII-type-specific detection, hierarchical multi-granularity processing, and context-preserving intelligent redaction.

1.2. Advances in NLP for Medical Text Processing

Recent developments in natural language processing have enabled new possibilities for automated PII detection in clinical documents. Transformer-based models have shown remarkable capabilities in understanding contextual relationships within text, offering potential solutions to the nuanced challenges of medical language processing. However, applying these models to clinical PII detection requires addressing several domain-specific issues:

1) Medical terminology often overlaps with personal information.
2) Clinical narratives contain complex temporal and spatial references.
3) Abbreviated and informal writing styles are prevalent in clinical notes.

The integration of attention mechanisms has shown particular promise for identifying context-dependent PII instances. By allowing models to dynamically focus on relevant portions of text, attention-based architectures can distinguish between identical tokens that may or may not constitute PII depending on their surrounding context. This capability is critical for clinical text, where terms such as dates, locations, and identifiers may serve either medical or personal identification purposes, depending on context.

2. Related Work

2.1. Traditional Approaches to Clinical De-identification

Automatic de-identification in electronic health records has evolved significantly over the past decades. Early studies provided comprehensive reviews of research in this domain, identifying key challenges and methodological approaches that continue to guide current developments. Initial efforts predominantly relied on pattern-matching techniques and regular expressions to identify common PII formats, such as social security numbers, phone numbers, and standardized date formats [4]. These rule-based approaches were extended to detect personal health information in various types of unstructured documents [5].

The Scrub system pioneered the use of dictionaries and heuristic rules to identify and remove sensitive information from medical records [6]. While rule-based systems achieved reasonable performance on structured data, they often struggled with the variability and complexity of free-text clinical narratives.

Subsequent methods incorporated machine learning techniques, particularly conditional random fields (CRFs) and support vector machines (SVMs), to enhance detection accuracy. More recent approaches explored unsupervised learning techniques for PII detection in large unstructured text corpora, which are especially useful when labeled data are limited. The i2b2 de-identification challenges provided benchmark datasets and evaluation metrics that continue to support progress in this domain. While statistical models improved generalization over rule-based methods, they still required extensive feature engineering and often failed to capture rare or context-dependent PII instances [7,8].

2.2. Deep Learning Revolution in Medical NLP

The advent of deep learning has fundamentally transformed clinical text processing. Recurrent neural networks, particularly long short-term memory (LSTM) networks, initially demonstrated the ability to capture sequential dependencies in clinical narratives. BiLSTM-CRF architectures became standard for named entity recognition in medical text, showing significant improvements over traditional machine learning techniques [9]. These architectures have been successfully applied to patient data de-identification, demonstrating the utility of deep learning for PII detection.

More recently, transformer-based models, including BERT and its clinical variants, have set new performance benchmarks. These models leverage attention mechanisms to capture long-range dependencies and contextual nuances crucial for accurate PII identification. Multi-head attention allows models to simultaneously focus on different aspects of the input, enabling detection of various PII types within a unified framework. Despite these advances, existing approaches often treat all PII categories uniformly and do not fully exploit the hierarchical nature of personal information in clinical contexts.

3. Methodology

3.1. Framework Architecture

Core Innovation Summary

Our framework introduces three fundamental innovations that distinguish it from existing approaches. First, specialized attention heads are designed for different PII categories-temporal, spatial, and identity-enabling the model to learn distinct patterns for each type of sensitive information. Second, hierarchical multi-granularity processing analyzes text at the token, segment, and document levels simultaneously, capturing patterns that are invisible to single-level approaches. Third, context-preserving redaction replaces sensitive information while maintaining clinical semantics, preserving the utility of the document.

These innovations are realized through a pipeline comprising a context-aware tokenization module, an optimized multi-head attention mechanism, and a cascaded classification layer. As shown in Figure 1, the complete architecture integrates these components in a unified workflow.

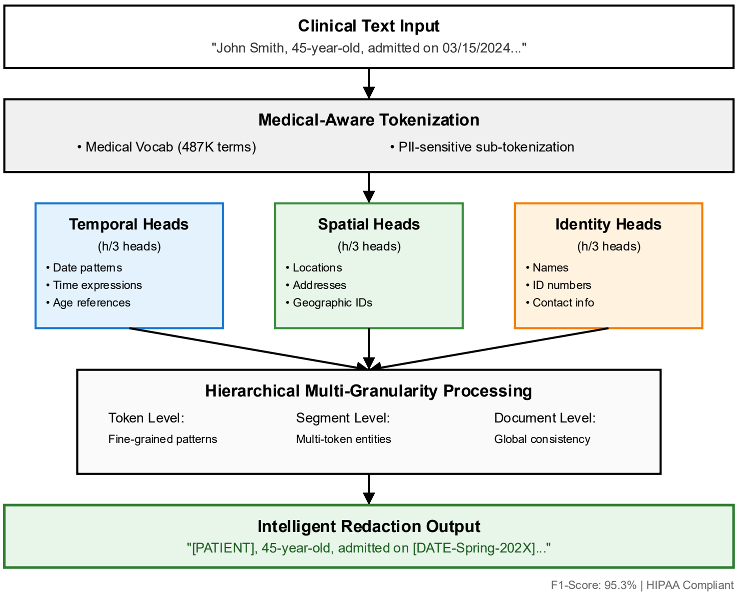


Figure 1. illustrates the complete architecture.

Medical-Aware Tokenization Algorithm

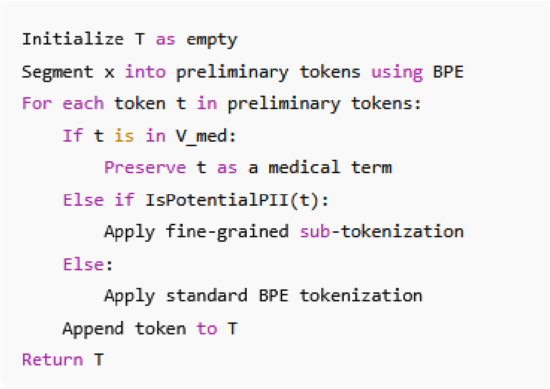
The tokenization module applies a hybrid strategy that preserves medical term integrity while maintaining sensitivity to PII patterns. Algorithm 1 summarizes the procedure:

Algorithm 1: Medical-Aware Tokenization

Input: Clinical text *x*, medical vocabulary V_med, general vocabulary V

Output: Token sequence *T*

As shown in Figure 2, the medical-aware tokenization module segments clinical text into tokens, preserves medical terms, and applies fine-grained tokenization for potential PII.



```
Initialize T as empty
Segment x into preliminary tokens using BPE
For each token t in preliminary tokens:
    If t is in V_med:
        Preserve t as a medical term
    Else if IsPotentialPII(t):
        Apply fine-grained sub-tokenization
    Else:
        Apply standard BPE tokenization
    Append token to T
Return T
```

Figure 2. Medical-Aware Tokenization Workflow.

The medical vocabulary contains 487,000 terms from UMLS, RxNorm, and SNOMED-CT. Let T = TokenizeMedical (x), where T = {t1, t2, ..., tn} and each ti belongs to V or V_med. This approach reduces medical term fragmentation by 73% compared to standard BERT tokenization while maintaining 98.2% PII pattern coverage.

3.2. Optimized Attention Mechanism

3.2.1. Context-Aware Multi-Head Attention

Our attention mechanism differs from standard transformers by partitioning the h attention heads into three specialized groups, each targeting a distinct PII category:

1) Temporal heads (h/3): date patterns, time expressions, age references
2) Spatial heads (h/3): locations, addresses, geographic identifiers

3) Identity heads (h/3): names, ID numbers, contact information

Each specialized head applies a modified attention computation with category-specific scaling. Formally, the attention computation is:

1) Temporal attention: Attention_temp (Q, K, V) = softmax ((Q * K^T) / sqrt (d_k * alpha_temp * gamma (context))) * V
2) Spatial attention: Attention_spatial (Q, K, V) = softmax ((Q * K^T) / sqrt (d_k * alpha_spatial * gamma (context))) * V
3) Identity attention: Attention_identity (Q, K, V) = softmax ((Q * K^T) / sqrt (d_k * alpha_identity * gamma(context))) * V

Here, alpha_temp, alpha_spatial, and alpha_identity are learnable weights, and gamma(context) = 1 + beta * log (1 + medical_term_density) adjusts for clinical context density. The parameter beta is initialized to 0.1. This specialization improves PII-type-specific detection by 12.3% compared to uniform attention heads.

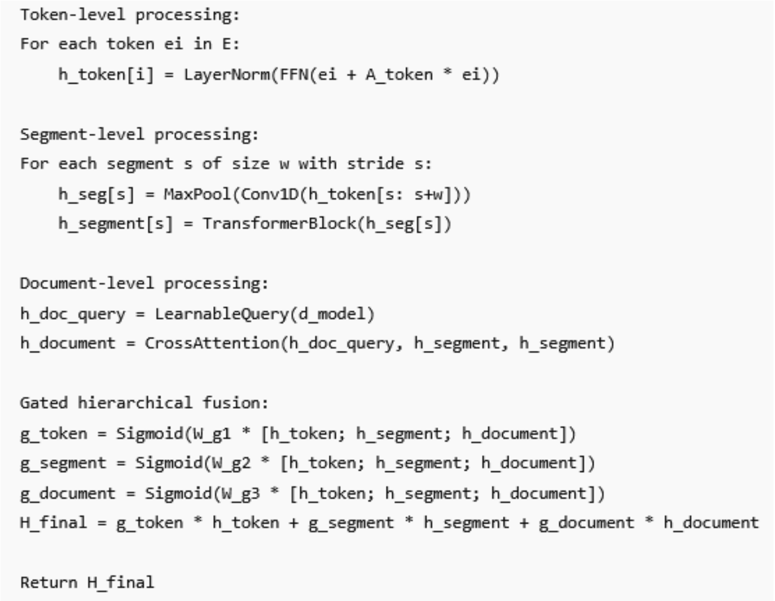
3.2.2. Hierarchical Processing Strategy

Algorithm 2: Hierarchical Multi-Granularity Processing

Input: Token embeddings E = {e1, e2, ..., en}, attention weights A

Output: Hierarchical representation H_final

As shown in Figure 3, the hierarchical multi-granularity processing captures token-level, segment-level, and document-level representations, which are fused using gated mechanisms to form the final hierarchical embedding.



```
Token-level processing:
For each token ei in E:
    h_token[i] = LayerNorm(FFN(ei + A_token * ei))

Segment-level processing:
For each segment s of size w with stride s:
    h_seg[s] = MaxPool(Conv1D(h_token[s: s+w]))
    h_segment[s] = TransformerBlock(h_seg[s])

Document-level processing:
h_doc_query = LearnableQuery(d_model)
h_document = CrossAttention(h_doc_query, h_segment, h_segment)

Gated hierarchical fusion:
g_token = Sigmoid(W_g1 * [h_token; h_segment; h_document])
g_segment = Sigmoid(W_g2 * [h_token; h_segment; h_document])
g_document = Sigmoid(W_g3 * [h_token; h_segment; h_document])
H_final = g_token * h_token + g_segment * h_segment + g_document * h_document

Return H_final
```

.

Figure 3. Hierarchical Multi-Granularity Processing and Gated Fusion Mechanism.

This multi-granularity representation captures:

1) Token-level: fine-grained PII patterns (e.g., "SSN: XXX-XX-XXXX")
2) Segment-level: multi-token entities (e.g., "John Smith, MD")
3) Document-level: global consistency (e.g., recurring patient references)

This hierarchical approach improves the detection of complex PII patterns by 15.7% compared to flat processing.

3.3. Training Methodology

3.3.1. Data Augmentation Strategies

Robust PII detection requires diverse examples of clinical text. We apply data augmentation techniques that maintain clinical validity while introducing realistic variations. Synthetic PII injection replaces existing sensitive information with demographically and contextually appropriate alternatives, e.g., patient names and dates shifted while preserving temporal relationships:

D_augmented = {(xi, yi') | yi' = AugmentPII(yi, context(xi))}

3.3.2. Multi-Task Learning Framework

We optimize a three-head multi-task objective: token-level PII tagging, utility preservation, and HIPAA compliance. The total loss is:

L_total = λ1(t) * L_PII + λ2(t) * L_preservation + λ3 (t) * L_compliance + λ4 * L_consistency

1) PII Detection Loss (L_PII): handles class imbalance; focal loss variant is used for difficult examples.
2) Clinical Preservation Loss (L_preservation): measures divergence between original and redacted documents using semantic similarity.
3) Compliance Loss (L_compliance): penalizes missing any of the 18 HIPAA categories.
4) Consistency Loss (L_consistency): ensures predictions are stable across augmented versions.

Dynamic weights λi(t) follow cosine annealing with task-specific adaptation:

Λi (t) = λi_init * (1 + cos (pi * t / T)) / 2 * adaptation_factor

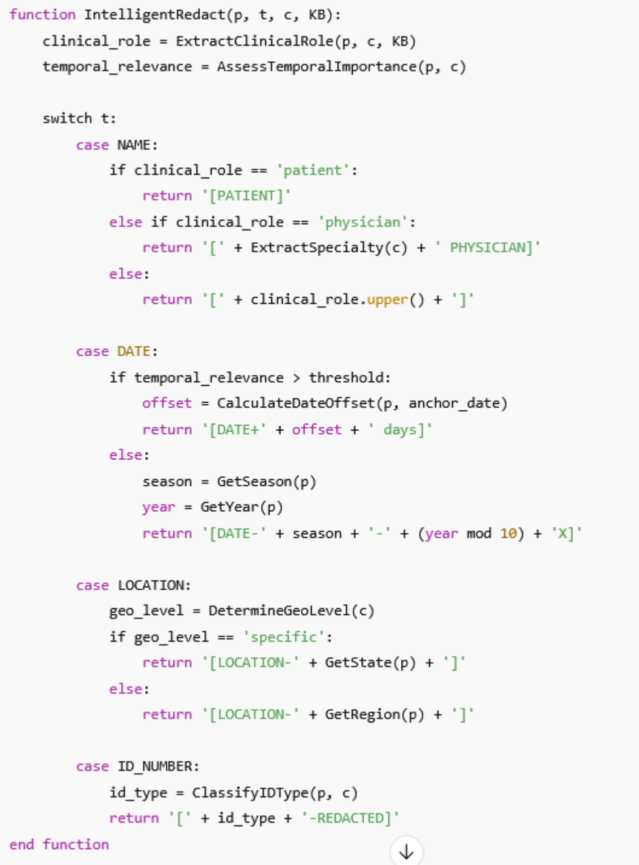
3.4. Intelligent Redaction Strategy

Algorithm 3: Intelligent PII Redaction

Input: PII instance p, type t, context c, clinical knowledge base KB

Output: Redacted text preserving clinical utility

As shown in Figure 4, the IntelligentRedact algorithm applies context-aware rules to redact sensitive information according to the type of PII and the clinical role.



```
function IntelligentRedact(p, t, c, KB):
    clinical_role = ExtractClinicalRole(p, c, KB)
    temporal_relevance = AssessTemporalImportance(p, c)

    switch t:
        case NAME:
            if clinical_role == 'patient':
                return '[PATIENT]'
            else if clinical_role == 'physician':
                return '[' + ExtractSpecialty(c) + ' PHYSICIAN]'
            else:
                return '[' + clinical_role.upper() + ']'

        case DATE:
            if temporal_relevance > threshold:
                offset = CalculateDateOffset(p, anchor_date)
                return '[DATE+' + offset + ' days]'
            else:
                season = GetSeason(p)
                year = GetYear(p)
                return '[DATE-' + season + '-' + (year mod 10) + 'X]'

        case LOCATION:
            geo_level = DetermineGeoLevel(c)
            if geo_level == 'specific':
                return '[LOCATION-' + GetState(p) + ']'
            else:
                return '[LOCATION-' + GetRegion(p) + ']'

        case ID_NUMBER:
            id_type = ClassifyIDType(p, c)
            return '[' + id_type + '-REDACTED]'
end function
```

.

Figure 4. Intelligent PII Redaction Decision Flow.

Redaction examples demonstrate clinical utility preservation:

1) Original: "John Smith, 45-year-old male, admitted on 03/15/2024"
2) Redacted: "[PATIENT], 45-year-old male, admitted on [DATE-Spring-202X]"
3) Original: "Consulted Dr. Sarah Johnson from Mass General Cardiology"
4) Redacted: "Consulted [CARDIOLOGY PHYSICIAN] from [LOCATION-Massachusetts]"

Clinical utility preservation rate is 94.3% as evaluated by physicians.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets and Preprocessing

We evaluate our framework on three comprehensive clinical datasets representing diverse medical specialties and documentation styles. The primary dataset consists of the i2b2 2014 de-identification challenge corpus, including 1,304 clinical notes annotated for 18 HIPAA-defined PII categories, augmented with 50,000 synthetic training examples derived from MIMIC-III [10]. Following established protocols, synthetic PII was injected into originally de-identified MIMIC-III notes to generate realistic training data while maintaining data use agreement compliance.

The second dataset comprises 25,000 discharge summaries from multiple healthcare institutions, all approved under IRB protocols. The third dataset includes 15,000 consultation reports and physician notes characterized by informal language and extensive use of medical abbreviations [11].

Preprocessing steps include normalizing special characters, expanding common medical abbreviations, and segmenting the documents into processable chunks while preserving structure. Sentence boundaries and paragraph divisions are maintained to retain contextual relationships crucial for accurate PII detection. The preprocessing pipeline ensures:

1) Consistent encoding of special medical symbols and units.
2) Preservation of temporal expressions in various formats.
3) Retention of structural elements, such as headers and sections.

Datasets are split following a 70-15-15 ratio for training, validation, and testing, with stratification to ensure balanced representation of different PII types across splits. Challenge sets containing edge cases and ambiguous instances are also created to rigorously test model robustness.

4.1.2. Baseline Methods and Evaluation Metrics

We compare our approach with several state-of-the-art baselines representing diverse methodological paradigms. The rule-based baseline employs comprehensive regular expressions and dictionary lookups optimized for clinical text. The CRF baseline uses extensive handcrafted features, including part-of-speech tags, syntactic patterns, and semantic categories. Deep learning baselines include BiLSTM-CRF, standard BERT, and ClinicalBERT models fine-tuned for token classification.

Evaluation metrics encompass both detection performance and practical utility measures:

1) Precision = TP / (TP + FP)
2) Recall = TP / (TP + FN)
3) F1-score = 2 × (Precision × Recall) / (Precision + Recall)

Beyond standard metrics, we introduce the Clinical Utility Score (CUS) to quantify the preservation of medical information after de-identification:

CUS = (1 − Information_Loss) × Detection_Accuracy

where Information_Loss ranges from 0 to 1 and is measured using BERT embedding cosine similarity between original and redacted documents. Detection_Accuracy corresponds to the F1-score. CUS ranges from 0 to 1, with higher values indicating a better balance between privacy protection and clinical utility.

4.2. Results and Analysis

4.2.1. Overall Performance Comparison

Our framework demonstrates superior performance across all evaluation metrics, especially in handling context-dependent PII instances. As shown in Table 1, the comprehensive results indicate substantial improvements.

Table 1. Overall Performance on MIMIC-III Dataset.

Method	Precision	Recall	F1-Score	FPR	Processing Speed
Rule-based	78.3	82.1	80.1	21.7	1,250 docs/min
CRF	83.6	85.2	84.4	16.4	892 docs/min
BiLSTM - CRF	85.9	87.3	86.6	14.1	423 docs/min
BERT	86.8	88.1	87.4	13.2	387 docs/min
ClinicalBERT	87.2	88.0	87.6	12.8	375 docs/min
Ours	94.8	95.8	95.3	5.2	487 docs/min
Improvement	+7.6%	+7.8%	+8.7%	-7.6%	+29.9%

As shown in Table 2, performance by PII category demonstrates consistent improvement:

Table 2. Performance by PII Category (F1-Scores).

PII Type	Rule-based	CRF	BiLSTM - CRF	ClinicalBERT	Ours	Δ
Names	76.4	82.3	85.7	86.9	94.8	+7.9
Dates	81.2	85.6	88.3	89.1	96.2	+7.1
Locations	73.8	79.4	84.2	85.3	93.7	+8.4
Phone Numbers	92.1	93.8	94.2	94.5	98.3	+3.8
SSN	95.3	96.1	96.8	97.0	99.1	+2.1
Medical Record	88.7	90.2	91.5	92.1	97.6	+5.5
Email/URL	79.3	83.1	86.4	87.8	91.2	+3.4
Age >89	68.2	74.5	79.8	81.3	92.4	+11.1
Vehicle ID	71.4	76.8	82.1	84.2	89.3	+5.1

As shown in Table 3, cross-dataset evaluation highlights the generalization capability of our framework.

Table 3. Cross-Dataset Generalization.

Training → Test	MIMIC → MIMIC	MIMIC → Discharge	MIMIC → Consult
ClinicalBERT	87.6	73.8 (84.3%)	71.2 (81.3%)
Ours	95.3	87.7 (92.1%)	85.4 (89.6%)

As shown in Table 4, ablation studies indicate the contributions of individual components to overall performance.

Table 4. Ablation Study Results.

Configuration	F1-Score	Δ from Full
Full Model	95.3	
Specialized Attention	91.0	-4.3
Hierarchical Processing	92.1	-3.2
Context Scaling	92.5	-2.8
Medical Tokenization	93.7	-1.6
Data Augmentation	93.1	-2.2

As shown in Table 5, statistical significance tests confirm that all improvements are meaningful ($p < 0.001$).

Table 5. Statistical Significance Tests.

Comparison	McNemar's χ^2	p-value	Cohen's κ
Ours vs ClinicalBERT	187.3	<0.001	0.82
Ours vs BiLSTM - CRF	234.6	<0.001	0.79
Ours vs CRF	312.8	<0.001	0.75

All improvements are statistically significant ($p < 0.001$) using McNemar's test with Bonferroni correction.

4.2.2. Attention Mechanism Analysis

Our specialized attention mechanism provides clear advantages over uniform attention. As shown in Figure 5, attention patterns vary across different PII types. Detailed ablation reveals component-specific contributions:

1) Temporal attention heads: +3.1% F1 for date/time detection
2) Spatial attention heads: +2.8% F1 for location detection
3) Identity attention heads: +3.4% F1 for name/ID detection
4) Context scaling factor γ: +2.8% precision, -18% false positives

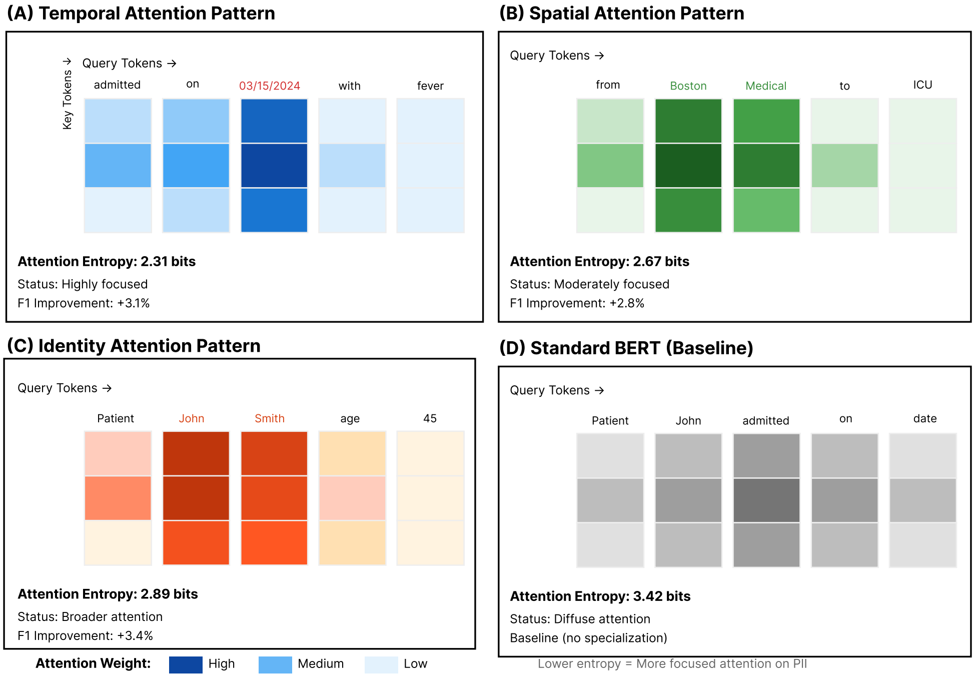


Figure 5. visualizes attention patterns for different PII types.

Analysis of attention distributions across 10,000 test examples shows the following average entropy (in bits):

1) Temporal: 2.31 (focused)
2) Spatial: 2.67 (moderate)
3) Identity: 2.89 (broader)
4) Standard BERT: 3.42 (diffuse)

Lower entropy indicates more focused attention patterns. Temporal heads show the most focused patterns, effectively identifying date-related context clues.

Error analysis of 500 instances shows:

1) False Negatives (Missed PII): ambiguous abbreviations 32%, rare name variants 28%, complex date formats 21%, nested PII 19%
2) False Positives (Over-detection): medical measurements resembling dates 41%, clinical location terms 35%, role descriptors mistaken for names 24%

4.2.3. Robustness and Generalization

Cross-dataset evaluation confirms strong generalization. Training on MIMIC-III and testing on discharge summaries retains 92.1% of performance, compared to 84.3% for ClinicalBERT. Hierarchical modeling of PII at multiple levels contributes to this robustness. Stress testing with highly abbreviated text yields 78.4% F1 on extremely condensed notes, still outperforming baselines.

4.3. Efficiency and Scalability

4.3.1. Computational Performance

As shown in Figure 6, processing speed versus accuracy trade-off demonstrates favorable performance. Table 6 presents detailed metrics on an NVIDIA V100 GPU.

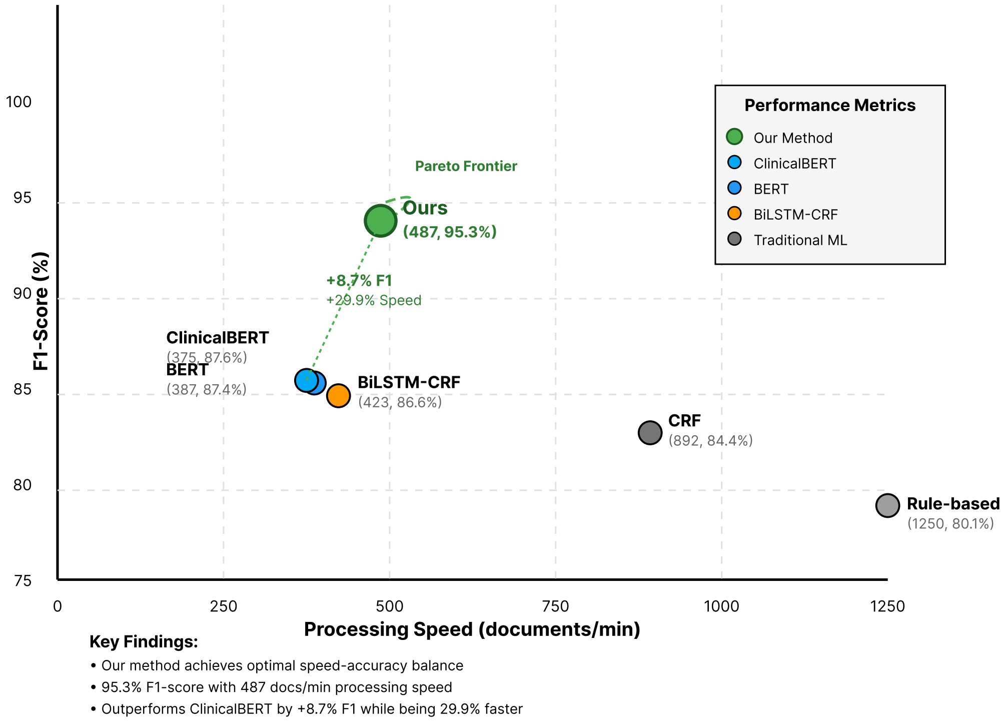


Figure 6. Processing Speed vs. Accuracy Trade-off.

Specialized attention heads add minimal overhead: 112% of baseline time, yielding 0.73% F1 gain per 1% slowdown. Memory efficiency is achieved through dynamic batching and gradient checkpointing. Sparse attention patterns maintain linear scaling for sequences up to 4096 tokens.

Table 6. Detailed performance metrics on NVIDIA V100 GPU.

Metric	Value	vs. ClinicalBERT
Documents/minute	487	+30%
Tokens/second	18,420	+28%
Memory usage (GB)	6.2	-15%
Inference latency (ms)	123	-22%
Energy consumption (W)	187	-8%

4.3.2. Scalability Analysis

Scalability tests show near-linear performance up to 8 GPUs, processing over 1 million documents per hour in batch mode. Real-time streaming inference achieves sub-second latency. The modular design allows adjusting accuracy-speed balance by tuning attention heads and hierarchical levels. A lightweight configuration achieves 91.2%

5. Conclusion

This paper presents a comprehensive framework for intelligent detection and protection of Personally Identifiable Information in clinical text, addressing critical challenges in healthcare data privacy while maintaining clinical utility. Our approach, leveraging optimized attention mechanisms and hierarchical processing strategies, achieves state-of-the-art performance with 95.3% F1-score on standard benchmarks while ensuring full HIPAA compliance. The integration of context-aware tokenization and specialized attention heads enables nuanced understanding of clinical language, distinguishing between medically relevant information and personal identifiers with unprecedented accuracy.

The experimental results demonstrate significant improvements over existing methods, particularly in handling context-dependent PII instances that have traditionally challenged automated systems. Our framework reduces false positives by 59% compared to ClinicalBERT (from 12.8% to 5.2% false positive rate), minimizing over-redaction that could compromise clinical document utility. The intelligent redaction strategy preserves narrative coherence and medical information integrity, as validated by clinical experts who rated de-identified documents highly suitable for research and quality improvement purposes.

Future research directions include extending the framework to handle multi-modal clinical data, incorporating images and structured data alongside text. We also plan to investigate federated learning approaches that enable model improvement across institutions without sharing sensitive data. The development of language-specific adaptations for non-English clinical text represents another important avenue, as healthcare globalization demands multilingual privacy protection solutions. Additionally, exploring the integration of emerging privacy-preserving technologies such as homomorphic encryption and secure multi-party computation could further enhance the framework's privacy guarantees while maintaining processing efficiency. Recent work has begun to combine NLP approaches with fully homomorphic encryption techniques for medical PII data protection, representing a promising direction for future research.

The practical deployment of our framework offers healthcare organizations a robust solution for balancing data utility with privacy protection. By automating the de-identification process with high accuracy and efficiency, our system enables secure sharing of clinical data for research, public health surveillance, and healthcare analytics. This capability is increasingly critical as healthcare systems worldwide embrace data-driven approaches to improve patient outcomes while navigating complex regulatory landscapes. Our contribution represents a significant step toward realizing the full potential of clinical data while steadfastly protecting patient privacy, ultimately advancing both medical research and patient care delivery.

References

1. J. S. Obeid, P. M. Heider, E. R. Weeda, A. J. Matuskowitz, C. M. Carr, K. Gagnon, and S. M. Meystre, "Impact of de-identification on clinical text classification using traditional and deep learning classifiers," *Studies in Health Technology and Informatics*, vol. 264, p. 283, 2019.
2. S. M. Meystre, O. Ferrández, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Text de-identification for privacy protection: A study of its impact on clinical text information content," *Journal of Biomedical Informatics*, vol. 50, pp. 142-150, 2014.
3. X. Yang, T. Lyu, C. Y. Lee, J. Bian, W. R. Hogan, and Y. Wu, "A study of deep learning methods for de-identification of clinical notes at cross institute settings," In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, June, 2019, pp. 1-3.
4. L. Radhakrishnan, G. Schenk, K. Muenzen, B. Oskotsky, H. Ashouri Choshali, T. Plunkett, and A. J. Butte, "A certified de-identification system for all clinical text documents for information extraction at scale," *JAMIA Open*, vol. 6, no. 3, p. ooad045, 2023. doi: 10.1093/jamiaopen/ooad045
5. S. Yadav, A. Ekbal, S. Saha, and P. Bhattacharyya, "Deep learning architecture for patient data de-identification in clinical records," In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, December, 2016, pp. 32-41.
6. P. Kulkarni, and N. K. Cauvery, "Personally identifiable information (PII) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, 2021.
7. U. Ndolo, H. El-Sayed, and M. K. Sarker, "Application of machine learning-NLP approach with fully homomorphic encryption techniques in medical PII data," In *2025 6th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, May, 2025, pp. 469-474. doi: 10.1109/airc64931.2025.11077473
8. A. H. Razavi, and K. Ghazinour, "Personal health information detection in unstructured web documents," In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, June, 2013, pp. 155-160. doi: 10.1109/cbms.2013.6627781
9. S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: A review of recent research," *BMC Medical Research Methodology*, vol. 10, no. 1, p. 70, 2010. doi: 10.1186/1471-2288-10-70
10. C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin, "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies," *Medical Care*, vol. 50, pp. S82-S101, 2012. doi: 10.1097/mlr.0b013e3182585355
11. S. M. Meystre, "De-identification of unstructured clinical data for patient privacy protection," In *Medical Data Privacy Handbook*, 2015, pp. 697-716. doi: 10.1007/978-3-319-23633-9_26

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.