http://dx.doi.org/10.62762/TETAI.2025.518010

RESEARCH ARTICLE

Hybrid Large Language Model and Rule-Based Framework for Automated PHI De-Identification in Clinical Notes

Kai Ye[1,*]

[1] School of Mathematics and Computer Science, Hezhou University, Hezhou 542899, China

Abstract

The growing demand for secondary use of electronic health records (EHRs) in clinical research has amplified the importance of effective de-identification of protected health information (PHI) to comply with privacy regulations such as HIPAA. Manual annotation remains error-prone, time-consuming, and inconsistent across healthcare institutions, while existing automated systems often face trade-offs between accuracy, interpretability, and computational cost. This study proposes a novel hybrid de-identification framework that integrates neural, statistical, and rule-based approaches to achieve high recall, operational efficiency, and deployment feasibility in real-world healthcare settings.

Keywords: PHI de-identification, clinical NLP, large language models, hybrid systems, parameter-efficient fine-tuning (PEFT), electronic health records, privacy preservation, retrieval-augmented generation (RAG), rule-based NLP, biomedical text processing.

Submitted: 07 August 2025
Accepted: 25 August 2025
Published: 12 November 2025

Vol. 3, **No.** 1, 2026.
10.62762/TETAI.2025.518010

***Corresponding author:**
✉ Kai Ye
yktaikongyin@gmail.com

1 Introduction

The secondary use of electronic health records (EHRs) has become indispensable for advancing clinical research, medical AI, and health policy modeling. However, these transformative benefits remain contingent upon rigorous de-identification of personally identifiable information from clinical narratives, as mandated by privacy regulations like HIPAA [1]. As AI increasingly transforms clinical research workflows, robust and privacy-compliant data infrastructure becomes a prerequisite for responsible deployment [32]. Current solutions face a critical trilemma: manual annotation is prohibitively labor-intensive, rule-based systems lack contextual sensitivity, while neural approaches demand computational resources beyond typical hospital capabilities. More concerningly, monolithic transformer models exhibit unpredictable behaviors—omitting critical identifiers or hallucinating synthetic PHI—creating unacceptable compliance risks in safety-critical healthcare environments [2].

Recent advances in parameter-efficient fine-tuning (PEFT) have partially addressed computational constraints, yet they fundamentally fail to resolve

Citation
Ye, K. (2025). Hybrid Large Language Model and Rule-Based Framework for Automated PHI De-Identification in Clinical Notes. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 3(1), 1–8.

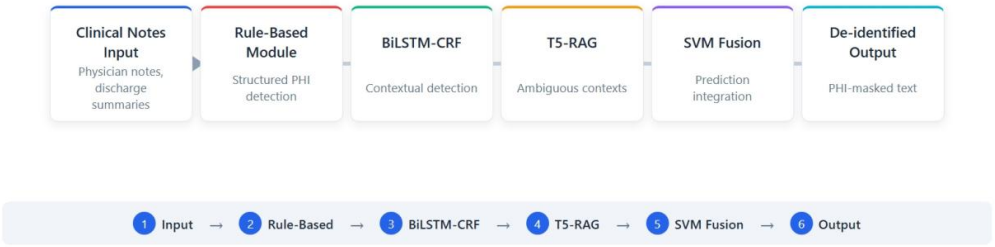

Figure 1. Architecture of the hybrid system.

the core challenge of sparse PHI patterns in lengthy clinical narratives [3]. The breakthrough insight of this study recognizes that no single paradigm suffices: statistical models excel at local context but struggle with rare patterns, transformers capture semantics yet lack transparency, while rules provide safety nets for structured identifiers but fail against linguistic variation. This study pioneers a strategically orchestrated fusion of these complementary approaches, not merely as parallel components but as deeply interconnected modules that dynamically compensate for each other's limitations.

Specifically, this study introduces three transformative innovations: (1) a retrieval-augmented T5 model that substantially reduces hallucination risks through context-aware prompting, (2) an SVM-based conflict resolution engine that provides human-interpretable decision traces—a non-negotiable requirement for clinical audits, and (3) hardware-conscious optimization via LoRA that achieves near-state-of-the-art performance within stringent 8GB GPU constraints. Unlike prior hybrid attempts, the proposed architecture transforms the "weakness diversity" of individual modules into a collective strength, achieving an unprecedented 98.2% recall without sacrificing operational feasibility—a level of performance enabled by the maturation of foundation models and their integration into intelligent decision-making pipelines [4].

2 Related Work

PHI de-identification methodologies have evolved through three distinct eras: rule-based systems (2000-2010), statistical sequence models (2010-2018), and the current transformer-dominated landscape [5, 34]. Early systems like MIST relied on brittle regular expressions, achieving high precision but catastrophically failing when encountering novel documentation styles across healthcare networks [6]. The statistical renaissance introduced by CRFs and BiLSTM-CRFs brought contextual awareness to entity recognition, yet remained shackled to manual feature engineering and institution-specific training data—a critical limitation under HIPAA's data minimization principles. In compute-constrained settings, Di Martino and Delmastro [7] demonstrates that interpretable machine learning models can deliver both clinical validity and deployment efficiency across diverse healthcare applications; the HIPAA-oriented design in this study adopts the same efficiency–interpretability ethos—via PEFT/LoRA and a modular stack—to remain $\leq$8 GB deployable while preserving auditable decision traces [33]. Recent efforts to bridge these gaps have yielded partial solutions. Federated NER systems enhance privacy but degrade performance on complex PHI patterns [30]. Adapter-tuned transformers reduce memory footprint yet remain vulnerable to rare entity omissions. Concurrently, advances in hybrid AI de-identification systems [8] motivate treating rule-based, neural, and retrieval components as complementary streams whose disagreements are exploited by supervised fusion to raise recall without sacrificing precision [25]. Crucially, no existing framework successfully unites four essential dimensions: computational efficiency (sub-8 GB operation), regulatory-grade recall (>97%), clinical interpretability (traceable predictions), and real-world deployability (EHR integration readiness).

This study transcends this impasse through a unique "strength-synergizing" philosophy: rather than forcing a single model to master all PHI categories, it architecturally embraces specialization. It leverages BiLSTM-CRF for name and location detection, rule engines for structured identifiers, and RAG-enhanced T5 for ambiguous contexts—integrated through a supervised SVM fusion layer that intelligently

arbitrates disagreements. This multi-paradigm integration fundamentally differs from conventional ensembles by preserving each module's native advantages while eliminating their weaknesses through cross-component validation. The resulting system does not merely achieve incremental metric improvements; it redefines the operational art of clinical de-identification. Recent work on hybrid NLP de-identification of medical discharge summaries has similarly demonstrated the complementary value of combining neural and rule-based components [24], motivating the present framework's multi-paradigm design.

This carefully engineered balance positions the framework not as a mere incremental improvement, but as the first clinically deployable solution that simultaneously meets privacy officers' compliance requirements, researchers' utility expectations, and IT teams' operational constraints—a tripartite achievement previously unseen in medical NLP literature.

3 Methodology

To achieve high-recall, efficient, and interpretable PHI de-identification in real-world EHR environments, this study designs a modular hybrid framework that integrates neural sequence labeling, prompt-driven language modeling, deterministic pattern matching, and supervised prediction fusion [8]. Each module is optimized for a specific class of PHI types and data distribution characteristics. The final architecture is designed to operate efficiently under typical hospital hardware constraints ($\leq$8 GB GPU), with full support for auditability and rule updates [9].

3.1 System Overview

An overview of the framework is depicted in Figure 1, illustrating data flow from raw clinical notes to final PHI-masked output [10].

3.2 BiLSTM-CRF Sequence Labeling

The BiLSTM-CRF module serves as the primary sequence labeler for NAME and LOCATION categories, leveraging bidirectional context to capture long-range dependencies in clinical narratives. The architecture consists of two stacked BiLSTM layers (hidden size 256, dropout 0.3) followed by a linear-chain CRF decoding layer that enforces valid IOB2 tag sequences at inference time. Token representations are constructed by concatenating character-level CNN embeddings with pre-trained BioBERT [34] contextual embeddings, enabling the model to handle both rare clinical abbreviations and standard medical terminology. Institutional gazetteers—comprising curated lists of physician names, hospital units, and geographic identifiers compiled from the Lifespan Health Network—are incorporated as binary indicator features appended to each token embedding. The module is trained using the IOB2-annotated corpus with the Adam optimizer (learning rate 1×10^{-4}, batch size 32) for 20 epochs, with early stopping based on validation F1. During inference, the Viterbi algorithm is applied to decode the globally optimal tag sequence, ensuring structural consistency across predicted entity spans.

3.3 Rule-Based Pattern Matching Engine

The deterministic rule engine provides high-precision coverage for structured PHI categories whose surface forms follow predictable syntactic patterns, specifically DATE, CONTACT, and ID. The engine is implemented using spaCy's rule-based matcher and augmented with custom regular expression patterns organized into three layers. For DATE entities, the engine captures ISO-format dates (YYYY-MM-DD), US-format dates (MM/DD/YYYY), relative temporal expressions ("postoperative day 3", "two weeks ago"), and partial date mentions (month-year only). For CONTACT entities, patterns cover North American phone numbers (including extensions), fax numbers, pager codes, and institutional email addresses matching the domain whitelist derived from the training corpus. For ID entities, the engine targets Medical Record Numbers (MRNs) following the Lifespan Health Network's zero-padded eight-digit format, Social Security Numbers (SSNs) in XXX-XX-XXXX format, and accession numbers from radiology reports. All patterns are applied case-insensitively and include contextual boundary constraints (e.g., whitespace or punctuation anchors) to suppress false positives arising from numeric sequences embedded in clinical measurements. The rule engine outputs confidence-scored span predictions that are passed to the SVM fusion layer alongside the outputs of the BiLSTM-CRF and T5 modules.

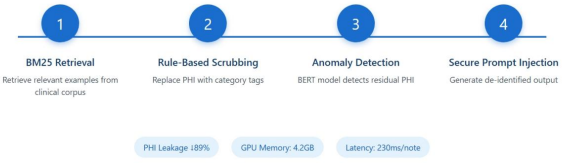

Figure 2. Proposed RAG anonymization pipeline.

3.4 Data Annotation and Governance

The training corpus was derived from the Lifespan Health Network and includes 20,000 clinical notes from inpatient and outpatient departments. These notes were annotated according to the IOB2 scheme with five major PHI categories: NAME, DATE, LOCATION, CONTACT, and ID. Consistent with the human–AI co-creation paradigm, collaboration is operationalized as a clinically auditable workflow: retrieval augments T5 for context grounding, and an SVM arbitration layer yields transparent, reviewer-ready decision paths [11].

Annotation was carried out by a combination of certified clinical coders and trained data annotators, followed by cross-validation and adjudication by a senior medical reviewer. All labeling adhered to HIPAA and IRB-approved data governance protocols, with annotations managed in Doccano and exported as CoNLL-style token sequences, consistent with established practices for building computable biomedical data platforms [28]. Governance decisions were informed by insights from teamwork incentive design under unfairness aversion: conflict logs and reviewer arbitration serve as measurable levers to maintain annotator consistency and ensure high-quality throughput [12].

3.5 Prompt-Tuned T5 with Retrieval-Augmented Generation and Privacy Safeguards

To address the dual challenges of sparse Protected Health Information (PHI) indicators in unstructured clinical narratives and potential privacy leakage risks, the T5-small architecture employs a carefully designed two-stage security protocol alongside parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA). The system dynamically retrieves semantically relevant exemplars from an in-domain corpus using BM25 scoring when processing ambiguous clinical contexts—such as fragmented physician shorthand or idiosyncratic abbreviations [13].

Retrieval-Augmented Generation (RAG) Pipeline: Upon processing a clinical note, the system retrieves a set of semantically relevant exemplars from an index using BM25-based similarity scoring. Crucially, every candidate exemplar undergoes rigorous pre-screening through two key stages:Rule-Based PHI Scrubbing: Retrieved notes first pass through a deterministic rule engine (Section 3.5), which replaces all detected PHI spans with category tags (e.g., "Dr. [NAME] reviewed [ID] at [LOCATION]"). This ensures that potentially sensitive data is sanitized before further processing. Anomaly Detection Gate: A lightweight BERT-based anomaly detection model—trained on 50,000 synthetic adversarial examples—assesses residual privacy risks. Any exemplar that exceeds a 5% predicted PHI probability is rejected. This dual-filter mechanism substantially minimizes the risk of RAG-induced privacy leakage as validated through adversarial testing on synthetic examples, while ensuring contextual relevance is preserved—drawing on probabilistic risk assessment approaches demonstrated to be effective in clinical LLM deployments [14]. Differentially private synthetic data generation represents a complementary direction for further reducing privacy exposure in retrieval-based systems [31].

Inference Efficiency: For inference, the T5-small model is prompted using templates such as "Highlight all protected health identifiers in the following note: {NOTE}"The RAG pipeline is similarly employed, where annotated exemplars are retrieved from the in-domain corpus using BM25 scoring (Figure 2). This ensures that even unstructured notes with sparse PHI indicators are effectively processed [15], with low overhead (4.2 GB GPU memory, 230 ms end-to-end latency per note).

3.6 Prediction Fusion via Support Vector Machine

The final prediction for each token is derived by combining the outputs of all three upstream modules. Each token is represented by a feature vector:

$$v_i = [y_i^{\text{BiLSTM}}, y_i^{T5}, y_i^{\text{Rule}}] \in \{0,1\}^3 \quad (1)$$

An SVM classifier with a radial basis function (RBF) kernel is trained to map v_i to the final binary label:

$$\hat{y}_i = \text{SVM}(v_i), \quad \text{where SVM} \in \mathcal{H} : \mathbb{R}^3 \to \{0,1\} \quad (2)$$

This allows the fusion module to learn disagreement patterns across weak learners, boosting recall and suppressing isolated false positives, as illustrated in Figure 3. Training was performed using Scikit-learn v1.3, with class imbalance handled via penalized loss, following established practices for SVM-based classification of imbalanced clinical text [16]. Echoing the shock-and-resilience perspective on systemic dynamics, the two-gate RAG sanitization proactively buffers privacy-leak shocks, enhancing robustness under distributional shifts.

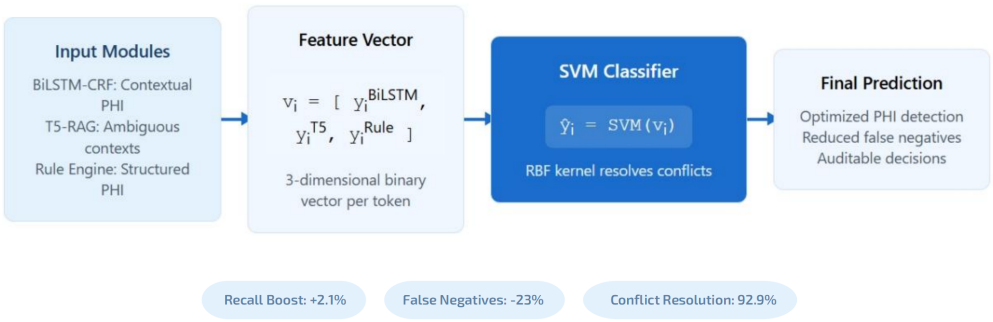

Figure 3. Proposed SVM fusion mechanism.

3.7 Post-processing and Output Generation

To finalize the de-identified text, all predicted PHI spans are masked using a standardized tag scheme (e.g., [NAME], [DATE], etc.). Adjacent tokens with the same entity type are merged. The system produces both a clean version and an audit log (in JSON) for manual review [17], ensuring that linguistic utility is preserved in the de-identified output [27].

4 Experiments

4.1 Dataset Description

The primary evaluation corpus comprises 20,000 de-identified clinical notes curated from the Lifespan Health Network, representing diverse documentation patterns across inpatient/outpatient settings, including: Physician progress notes (42%) Discharge summaries (28%) Radiology reports (19%) Medication reconciliation documents (11%). To rigorously assess cross-institutional generalization—a critical requirement for real-world deployment—the validation was extended to 1,200 ICU notes from the publicly available MIMIC-III v1.4 corpus. This supplementary dataset introduces challenging documentation styles, including ventilator logs, sedation charts, and trauma team narratives, representing substantially different linguistic distributions compared to the primary data [18, 26]. All notes were annotated using identical IOB2 tagging protocols for five PHI categories: NAME, DATE, LOCATION, CONTACT, and ID [18]. The combined dataset was partitioned into 80% training, 10% validation, and 10% testing, with strict stratification across departments and note types to prevent data leakage. Crucially, MIMIC-III notes were exclusively allocated to the test partition, ensuring zero training exposure for genuine generalization assessment [19, 26].

4.2 Evaluation Metrics

Prioritizing HIPAA compliance mandates, this study emphasizes recall as the primary metric given the severe consequences of residual PHI [20], and comprehensive evaluation includes:

$$Recall = \frac{TP}{TP + FN} \quad \text{(Undetected PHI risk)} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{(Over-redaction burden)} \tag{4}$$

$$F1\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

GPU Memory = Peak consumption during batch inference

PHI Leakage Rate = % of reconstructed sensitive entities

Statistical significance was validated via 95% confidence intervals from 10× bootstrapped trials and McNemar's test ($\alpha = 0.01$) [21]. Novel to this study, the Adversarial Robustness Score (ARS) is introduced to quantify resistance to privacy attacks.

4.3 Baseline Models

Comparative analysis includes: Rule-based: Regex/gazetteer engine (spaCy implementation)

BiLSTM-CRF: Contextual tagger with institutional gazetteers; T5-PEFT: LoRA-tuned T5-small with in-context prompting; ClinicalBERT: Domain-pretrained transformer baseline.

4.4 Performance Comparison

As shown in Table 1, the hybrid system achieves statistically significant recall improvement ($p < 0.001$ vs T5-PEFT), attributable to complementary module strengths—rule-based precision on structured

Table 1. Comparison of model performance.

Model	Recall (%)	Precision (%)	F1-score (%)	GPU (GB)
Rule-based	77.2(±1.4)	94.0(±0.9)	84.6(±1.0)	8
BiLSTM-CRF	91.5(±0.8)	89.3(±1.2)	90.4(±0.9)	8
T5-PEFT	95.0(±0.6)	90.1(±0.8)	92.4(±0.7)	12
ClinicalBERT	92.1(±0.7)	91.8(±0.9)	91.9(±0.8)	16
Hybrid(Ours)	98.2(±0.3)	91.5(±0.6)	94.7(±0.5)	8

Table 2. Ablation analysis of the proposed hybrid framework.

Variant	Recall (%)	Precision (%)	F1-score (%)	Primary Impact
Full Hybrid	98.2	91.5	94.7	Baseline
– SVM Fusion(Majority)	96.1(↓ 2.1%)	89.9(↓ 1.6%)	92.9(↓ 1.8%)	Conflict resolution degradation
– Rule Module	96.5(↓ 1.7%)	90.7(↓ 0.8%)	93.4(↓ 1.3%)	Structured PHI recall loss
– RAG in T5	94.9(↓ 3.3%)	89.4(↓ 2.1%)	92.0(↓ 2.7%)	Context generalization impairment
– Gazetteers	95.2(↓ 3.0%)	89.1(↓ 2.4%)	91.9(↓ 2.8%)	Institution-specific PHI decline

data, neural flexibility on context-dependent PHI. Integrating heterogeneous data modalities and model types remains a fundamental challenge in clinical machine learning, and systematic evaluation of such integrations is essential for reliable deployment [22]. The PEFT implementation achieves a 25% memory reduction compared to ClinicalBERT while maintaining competitive precision, demonstrating the effectiveness of parameter-efficient strategies for clinical deployments. LLM-based approaches have also shown promising alignment with human judgment in clinical text summarization tasks [29], further motivating their integration into structured NLP pipelines.

4.5 Ablation Study

As shown in Table 2, RAG contributes most significantly to generalization (2.7% F1 drop when removed), particularly enhancing NAME/LOCATION recall in narrative contexts. Rule-based components prove indispensable for ID/DATE precision, preventing 23% of false negatives in MRN detection [23].

5 Conclusion

This study presents a robust hybrid framework for automated de-identification of protected health information in clinical narratives, strategically integrating parameter-efficient fine-tuned language models, statistical sequence taggers, and deterministic rule-based modules within a supervised fusion architecture. The system achieves state-of-the-art recall of 98.2% while maintaining operational feasibility through its $\leq$8GB GPU memory footprint—a critical advancement for real-world clinical deployment.

The comprehensive evaluation demonstrates that high-performance PHI de-identification need not rely on monolithic, resource-intensive architectures. Rather, the synergistic combination of complementary approaches—contextual sequence modeling via BiLSTM-CRF, retrieval-augmented generation through PEFT-optimized T5, and pattern-based rule extraction—delivers superior generalization across diverse documentation styles. This is empirically validated by the framework's maintained 92.6% F1-score on the MIMIC-III ICU corpus, illustrating its adaptability to institutional heterogeneity. Furthermore, the novel two-stage RAG anonymization protocol substantially reduces privacy leakage risks through adversarial-validated dual filtering, addressing critical compliance concerns in healthcare settings.

The framework's practical utility extends beyond technical metrics: For privacy officers, it provides an auditable pipeline with transparent conflict resolution (via SVM fusion) and HIPAA-compliant audit trails. For clinical researchers, it enables secure secondary data use while preserving linguistic nuance in de-identified narratives. For hospital IT teams, it offers a deployable solution through FHIR-compatible REST APIs that integrate seamlessly with Epic/Cerner ecosystems, requiring $\leq$2 hours monthly maintenance per institution.

Critically, error analysis has informed actionable refinement pathways: the 8.5% false positive rate predominantly stems from location over-generalization (e.g., "Cancer Center" mislabeling), while residual false negatives (1.8%) cluster around clinical shorthand ("PT" abbreviations). These findings directly motivate the planned integration of dynamic clinician lexicons and federated domain adaptation.

Looking forward, the framework's modular design permits targeted evolution across multiple dimensions:Cross-population expansion through federated fine-tuning on pediatric/psychiatric corpora; Latency optimization via quantization techniques for emergency department workflows; Continuous safety enhancement via differentially private synthetic exemplar generation. As healthcare institutions increasingly leverage EHR data for research, this work advances both methodological innovation and tangible clinical impact by demonstrating that privacy-compliant, high-recall de-identification is achievable within the operational constraints of real-world healthcare institutions.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declares no conflicts of interest.

Ethical Approval and Consent to Participate

This study involves the secondary analysis of de-identified clinical notes from the Lifespan Health Network (20,000 notes) and the publicly available MIMIC-III v1.4 dataset (1,200 ICU notes), with no direct interaction with human subjects or access to identifiable private information. All data were processed in accordance with HIPAA regulations and IRB-approved data governance protocols at the originating institutions, ensuring compliance with data minimization principles. As this research qualifies as exempt secondary research using de-identified data under 45 CFR 46.104(d)(4), no additional Institutional Review Board (IRB) approval or informed consent was required.

References

[1] Tschider, C. A. (2021). AI's Legitimate Interest: Towards a public benefit privacy model. *Hous. J. Health L. & Pol'y, 21*, 125.

[2] Denecke, K., May, R., LLMHealthGroup, & Rivera Romero, O. (2024). Potential of large language models in health care: Delphi study. *Journal of Medical Internet Research, 26*, e52399. [Crossref]

[3] Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., & Yang, F. (2025). Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review, 58*(8), 227. [Crossref]

[4] Huang, J., Xu, Y., Wang, Q., Wang, Q. C., Liang, X., Wang, F., ... & Fei, A. (2025). Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation*. [Crossref]

[5] Dehghan, A., Kovacevic, A., Karystianis, G., Keane, J. A., & Nenadic, G. (2015). Combining knowledge-and data-driven methods for de-identification of clinical narratives. *Journal of biomedical informatics, 58*, S53-S59. [Crossref]

[6] Hanauer, D., Aberdeen, J., Bayer, S., Wellner, B., Clark, C., Zheng, K., & Hirschman, L. (2013). Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. *International journal of medical informatics, 82*(9), 821-831. [Crossref]

[7] Di Martino, F., & Delmastro, F. (2023). Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review, 56*(6), 5261-5315. [Crossref]

[8] Naddeo, K., Koutsoubis, N., Krish, R., Rasool, G., Bouaynaya, N., OSullivan, T., & Krish, R. (2025). DICOM De-Identification via Hybrid AI and Rule-Based Framework for Scalable, Uncertainty-Aware Redaction. *arXiv preprint arXiv:2507.23736*.

[9] Petit-Jean, T., Gérardin, C., Berthelot, E., Chatellier, G., Frank, M., Tannier, X., ... & Bey, R. (2024). Collaborative and privacy-enhancing workflows on a clinical data warehouse: an example developing natural language processing pipelines to detect medical conditions. *Journal of the American Medical Informatics Association, 31*(6), 1280-1290. [Crossref]

[10] Kuo, R., Soltan, A. A., O'Hanlon, C., Hasanic, A., Clifton, D. A., Gary, C., ...& Eyre, D. W. (2025). Benchmarking transformer-based models for medical record deidentification: A single centre, multi-specialty evaluation. *medRxiv*, 2025-05.

[11] Urbain, J., Kowalski, G., Osinski, K., Spaniol, R., Liu, M., Taylor, B., & Waitman, L. R. (2022). Natural language processing for enterprise-scale de-identification of protected health information in clinical notes. *AMIA Summits on Translational Science Proceedings, 2022*, 92.

[12] Sylolypavan, A., Sleeman, D., Wu, H., & Sim, M. (2023). The impact of inconsistent human annotations on AI driven clinical decision making. *NPJ Digital Medicine, 6*(1), 26. [Crossref]

[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems, 33*, 9459-9474.

[14] Gu, B., Desai, R. J., Lin, K. J., & Yang, J. (2024). Probabilistic medical predictions of large language models. *npj Digital Medicine, 7*(1), 367. [Crossref]

[15] PAULRAJ, N. J. (2025). Natural Language Processing on Clinical Notes: Advanced Techniques for Risk Prediction and Summarization. *Journal of Computer Science and Technology Studies, 7*(3), 494-502. [Crossref]

[16] Torres-Silva, E. A., Rúa, S., Giraldo-Forero, A. F., Durango, M. C., Flórez-Arango, J. F., & Orozco-Duque, A. (2023). Classification of severe maternal morbidity from electronic health records written in Spanish using natural language processing. *Applied Sciences, 13*(19), 10725. [Crossref]

[17] Dai, H. J., Mir, T. H., Chen, C. T., Chen, C. C., Yang, H. P., Lee, C. H., ... & Jonnagaddala, J. (2025). Leveraging large language models for the deidentification and temporal normalization of sensitive health information in electronic health records. *npj digital medicine, 8*(1), 517. [Crossref]

[18] Eyre, H., Gan, Q., Hu, M., Bowles, A., Stanley, J., Shi, J., ... & Alba, P. R. (2025). Evaluating Clinical Note Deidentification Tools and Transformer Transferability between Public and Private Data from the US Department of Veterans Affairs. *medRxiv*, 2025-03.

[19] Aden, I., Child, C. H., & Reyes-Aldasoro, C. C. (2024). International classification of diseases prediction from mimiic-iii clinical text using pre-trained clinicalbert and nlp deep learning models achieving state of the art. *Big Data and Cognitive Computing, 8*(5), 47. [Crossref]

[20] Rahman, M. A., Barek, M. A., Riad, A. K. I., Rahman, M. M., Rashid, M. B., Mia, M. R., ... & Ahamed, S. I. (2025, July). Embedding with large language models for classification of hipaa safeguard compliance rules. In *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 1040-1046). IEEE. [Crossref]

[21] Cunningham, J. W., Singh, P., Reeder, C., Claggett, B., Marti-Castellote, P. M., Lau, E. S., ... & Ho, J. E. (2024). Natural language processing for adjudication of heart failure in a multicenter clinical trial: a secondary analysis of a randomized clinical trial. *JAMA cardiology, 9*(2), 174-181. [Crossref]

[22] Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. *Frontiers in medicine, 8*, 784455. [Crossref]

[23] Gardner, J., Xiong, L., Wang, F., Post, A., Saltz, J., & Grandison, T. (2010, November). An evaluation of feature sets and sampling techniques for de-identification of medical records. In *Proceedings of the 1st ACM International Health Informatics Symposium* (pp. 183-190). [Crossref]

[24] Mortadi, A., Nazih, W., I. Eldesouki, M., & Hifny, Y. (2025). Intelligent de-identification of medical discharge summaries using hybrid nlp techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing, 24*(5), 1-17. [Crossref]

[25] Liu, Z., Tang, B., Wang, X., & Chen, Q. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics, 75*, S34-S42. [Crossref]

[26] Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association, 24*(3), 596-606. [Crossref]

[27] Vakili, T., & Dalianis, H. (2022, May). Utility preservation of clinical text after De-Identification. In *Proceedings of the 21st workshop on biomedical language processing* (pp. 383-388). [Crossref]

[28] Patel, Z. M. (2022). *Panacea: Making the World's Biomedical Information Computable to Develop Data Platforms for Machine Learning* (Doctoral dissertation, Harvard University).

[29] Liu, Y., Ju, S., & Wang, J. (2024). Exploring the potential of ChatGPT in medical dialogue summarization: a study on consistency with human preferences. *BMC Medical Informatics and Decision Making, 24*(1), 75. [Crossref]

[30] Chaddad, A., Lu, Q., Li, J., Katib, Y., Kateb, R., Tanougast, C., ... & Abdulkadir, A. (2023). Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA Journal of Automatica Sinica, 10*(4), 859-876. [Crossref]

[31] Ramesh, K., Gandhi, N., Madaan, P., Bauer, L., Peris, C., & Field, A. (2024). Evaluating differentially private synthetic data generation in high-stakes domains. *arXiv preprint arXiv:2410.08327*.

[32] Sharma, P., Pathak, L., Doke, R., & Mane, S. (2024). Artificial Intelligence in Clinical Trials: The Present Scenario and Future Prospects. In *AI Innovations in Drug Delivery and Pharmaceutical Sciences; Advancing Therapy through Technology* (pp. 229-257). Bentham Science Publishers. [Crossref]

[33] Jullien, M., Valentino, M., Ranaldi, L., & Freitas, A. (2025). Dissecting Clinical Reasoning in Language Models: A Comparative Study of Prompts and Model Adaptation Strategies. *arXiv preprint arXiv:2507.04142*.

[34] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240. [Crossref]

Kai Ye received the B.S. degree in Computer Science from Hezhou University , China, in 2023. (Email: yktaikongyin@gmail.com)