Transactions on Computational and Scientific Methods | Vo. 4, No. 08, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Adaptive Privacy-Aware Federated Language Modeling for Collaborative Electronic Medical Record Analysis

Anzhuo Xie
Columbia University, New York, USA
xieanzhuo@outlook.com

Abstract: This study addresses the challenges of non-shareable data, significant semantic variation across institutions, and strict privacy constraints in collaborative electronic medical record settings. It proposes a federated language modeling framework for electronic medical records and introduces an adaptive privacy budget scheduling algorithm to improve model stability and applicability in real medical environments. The method builds a local medical text encoding module at each institution to convert raw records into continuous semantic representations and uses semantic factorization to separate latent representations into generalizable causal semantic factors and sensitive factors that require protection. This enables explicit distinction between key semantic structures and private information. During federated training, the framework constructs a unified semantic space through cross-institution semantic alignment and adjusts noise injection dynamically through the adaptive privacy budget mechanism to balance privacy protection and semantic usability. To evaluate its effectiveness, the study includes multiple comparative experiments and sensitivity analyses, examining performance, budget scheduling strategies, and variations in training conditions. The results show that the framework maintains strong semantic representation under strict privacy constraints and outperforms several baseline models across multiple metrics, demonstrating the necessity and effectiveness of building semantically decomposable and privacy-adaptive federated language models for cross-institution electronic medical record tasks. Overall, the proposed method provides a feasible solution for high-quality medical text modeling under privacy-restricted conditions and shows strong potential for multi-center medical data collaboration.

Keywords: Federated language modeling; electronic medical record text; privacy budget allocation; semantic factorization

1. Introduction

The rapid accumulation of electronic medical records in recent years has driven the development of medical natural language processing[1]. As the core clinical text resource in healthcare institutions, electronic medical records contain key information related to disease progression, treatment strategies, examination indicators, and risk alerts. They form the foundation for clinical decision support and intelligent analysis. However, these records combine structured and unstructured content. They include professional terminology, abbreviations, nested semantics, and sensitive information linked to patient identity. This produces major challenges for traditional centralized modeling in data collection, cross-institution data sharing, and model generalization. In addition, the writing style, data quality, departmental distribution, and workflow vary significantly across institutions. These differences create strong data heterogeneity that reduces the stability and reliability of models across scenarios. Under the demand for collaborative use of multi-source data, high-

quality language modeling of electronic medical records with strict privacy protection has become an important topic in medical artificial intelligence[2].

With stricter privacy regulations and increasing compliance requirements in healthcare, it is necessary to build a learning framework that protects sensitive information while enabling cross-institution collaboration. Federated learning provides a new paradigm that allows institutions to train models without centralizing data. It reduces the risks of data silos and privacy leakage. However, classical federated learning methods are primarily designed for structured features or images. Electronic medical records are long texts with sparse semantics and domain-specific language features. Their representations and distributions are more complex. Semantic shifts, differences in document segmentation practices, inconsistent terminology systems, and variability in text-generation workflows across institutions amplify the client drift problem in federated modeling. Therefore, federated learning for electronic medical records requires a language framework that unifies medical text semantics and remains robust in heterogeneous text environments. This ensures convergence, semantic consistency, and reliable generalization during cross-institution collaboration[3].

In multi-institution collaboration, the allocation and control of privacy budgets become central to the security and usability of federated systems. Differential privacy is the most widely accepted mechanism with theoretical guarantees. It injects noise into parameters or gradients to limit information leakage. Its protection strength depends on proper budget allocation. Fixed or coarse-grained budgets cannot adapt to dynamic semantic gradients, differences in data scales across institutions, or cumulative privacy loss during repeated communication. Improper allocation produces risks. When the budget is too tight, excessive noise disrupts semantic features and weakens model convergence and language understanding. When the budget is too loose, privacy protection weakens, and sensitive information may be exposed during training. A mechanism that adapts privacy budgets to task structure, semantic complexity, and client state is therefore essential for a federated language system designed for electronic medical records.

These challenges call for a new federated language framework that unifies cross-institution text representations while balancing semantic consistency and privacy constraints. Such a framework must learn a stable medical semantic space without direct data sharing. It must support long-text modeling, terminology disambiguation, semantic normalization across institutions, and robust contextual understanding. As collaboration scales up, corpus heterogeneity increases. Cross-institution semantic alignment and knowledge transfer without exposing raw text has become essential for building usable, controllable, and trustworthy medical language models. Integrating federated learning with medical text modeling and incorporating adjustable privacy budget management can provide a secure and controllable pathway for intelligent analysis of electronic medical records. It also lays the foundation for building a large-scale collaborative ecosystem of medical text[4,5].

From a broader perspective, research on federated language frameworks and adaptive privacy budget mechanisms for electronic medical records has both theoretical and practical value[6]. It offers a viable approach for healthcare institutions to collaborate under strict privacy constraints. It helps break data silos and enables high-quality multi-center text knowledge sharing. It also improves the deployability of medical natural language processing models in real clinical settings, supporting decision assistance, clinical reasoning, health management, and risk prediction. Furthermore, this research promotes deeper integration of privacy protection theory, federated optimization, and domain-specific language modeling[7]. It provides a more robust, secure, and scalable foundation for future medical artificial intelligence systems. Therefore, exploring a federated language framework for electronic medical records and designing an adaptive privacy budget mechanism are important for strengthening data utilization, enhancing cross-institution collaboration, and advancing the coordinated development of privacy protection and artificial intelligence.

2. Proposed Framework

To achieve secure collaborative modeling of electronic medical records across institutions, this study constructs a federated language framework for long medical texts. Through four core components-local text encoding, semantic factor decomposition, cross-institutional alignment, and privacy control generalizable unified medical semantic space is formed. First, each participating institution constructs a domain-specific contextual representation based on its local electronic medical record text and maps the original text into continuous latent semantic vectors, thus completing language modeling without exposing the original content. This paper also presents the overall model architecture diagram, and its experimental results are shown in Figure 1.

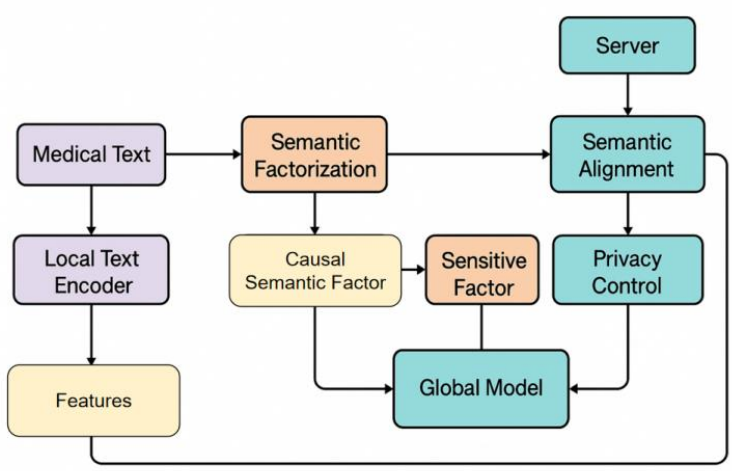

Figure 1. Overall model architecture diagram

Let institution k's local corpus be x_k, and the latent semantic representation generated by the local language model be:

$$h_k = f_{\theta_k}(X_k)$$

Where f_{θ_k} is an independently trainable medical text encoding model. This implicit representation avoids direct text transmission, laying the foundation for subsequent cross-institutional semantic fusion.

To further remove unnecessary sensitive factors related to patient identity, the framework introduces a causal semantic-privacy bi-subspace decomposition mechanism, which decomposes the text representation into a generalizable causal semantic part and a sensitive feature part that needs protection. Let the latent space representation be h_k, which is obtained through decomposition mapping:

$$h_k = h_k^{(C)} + h_k^{(S)}$$

Where $h_k^{(C)}$ represents structurally stable causal semantic factors, and $h_k^{(S)}$ represents sensitive factors related to individual-specific records, institutional writing biases, etc. During the optimization process, structural consistency constraints are incorporated to ensure the causal semantic space remains alignable across multiple institutions. This consistency constraint can be defined as:

$$L_{align} = \sum_{i \neq j} \| h_i^{(c)} - h_j^{(c)} \|_2^2$$

This enables institutions to form a unified causal semantic structure without sharing text.

During the cross-institutional collaborative training phase, the framework employs a secure aggregation mechanism to integrate the causal semantic gradients from each client. Simultaneously, it designs an adaptive privacy budget adjustment algorithm oriented towards text features to address the issue of gradient sensitivity varying with semantic complexity. Specifically, each client applies differential privacy perturbation to the gradients before uploading, with the perturbation taking the following form:

$$\widetilde{g}_k = g_k + N(0, \sigma_k^2 I)$$

Where g_k is the local gradient, and σ_k^2 is dynamically determined based on an adaptive privacy budget strategy. Privacy loss is accumulated according to the fundamental combination property of differential privacy, satisfying:

$$\varepsilon = \sum_{t=1}^{T} \varepsilon_t$$

Where ε_t is the privacy budget for the t-th round of communication. The adaptive mechanism dynamically adjusts the noise level based on the gradient norm, semantic distribution drift, and communication round, achieving a balance between privacy protection and semantic availability.

During the global model update phase, the system performs weighted fusion based on the causal semantic contribution and budget consumption of each client to construct a stable global medical semantic model. Let $\widetilde{g}_k$ be the perturbation gradient uploaded by each institution, and its global aggregation form be:

$$g_{global} = \sum_{k=1}^{K} \alpha_t \widetilde{g}_k$$

The weight α_t considers corpus size, semantic consistency, and privacy budget status simultaneously. This mechanism ensures that the framework can still form a semantically stable and cross-institutional generalizable medical language space while maintaining strong privacy constraints. Through a closed-loop structure of text encoding-semantic decomposition, control-global alignment, this method achieves secure, efficient, and structured federated language modeling for the sensitive text scenario of electronic medical records.

3. Experimental Analysis

3.1 Dataset

This study uses the eICU Collaborative Research Database as the primary source of electronic medical records. The dataset is constructed by multiple intensive care institutions. It contains clinical notes, nursing documentation, condition observations, descriptions of diagnostic and treatment procedures, and both structured and unstructured texts. Its multi-institution, multi-style, and multi-scenario characteristics lead to clear differences in writing styles, terminology usage, and documentation standards. These features reflect the heterogeneity of cross-institution electronic medical records. The dataset, therefore, provides an ideal basis for studying unified semantic modeling, cross-institution semantic alignment, and privacy protection mechanisms under federated conditions.

Because the eICU data originate from different healthcare institutions, the structure, paragraph organization, and writing templates vary widely. The records differ in length, terminology systems, and contextual dependencies. These characteristics are typical of clinical texts. They place higher demands on federated language modeling. The model must achieve cross-institution robustness, semantic generalization, and style independence. In addition, compared with open-domain text tasks, electronic medical records contain large amounts of sensitive information, such as descriptions of diseases, treatment actions, and clinical pathways.

The data cannot be shared directly across institutions. This provides a realistic setting for evaluating adaptive privacy budget strategies and secure aggregation mechanisms.

In this study, the data are divided into multiple independent clients according to the assumptions of federated learning. Each client represents a different nursing institution or medical unit. Text encoding and semantic factor extraction are performed only at the local site. Only the unstructured text components are used. These include admission notes, nursing notes, condition observations, and treatment summaries. They serve as the inputs for federated semantic modeling and provide continuous, high-information clinical language resources for building a unified semantic space across institutions. With the multi-center characteristics of the eICU dataset, this study can evaluate the applicability of the federated language framework and adaptive privacy budget mechanism under strict privacy constraints in a real medical environment.

3.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	Acc	Precision	Recall	AUC
XGBoost[8]	0.842	0.828	0.811	0.873
Transformer[9]	0.867	0.854	0.842	0.892
BILSTM[10]	0.853	0.847	0.821	0.885
BERT[11]	0.881	0.872	0.859	0.908
Ours	0.907	0.894	0.883	0.936

The comparative results show that traditional machine learning models face clear limitations when processing cross-institution electronic medical records. XGBoost performs well on structured features. However, it cannot capture long-range context in long texts or records with diverse writing styles. It also struggles with wide semantic spans in medical documentation. As a result, it produces the lowest scores across all metrics. BiLSTM and Transformer outperform traditional models. They can model semantic dependencies and local contextual structures to some extent. Yet they remain sensitive to variations in writing style, inconsistent terminology, and noise introduced by privacy protection. Their overall performance, therefore, does not reach the optimal level.

Pretrained language models show stronger capability in medical text processing. BERT achieves high results in precision, recall, and AUC. This indicates that its deep semantic representations help reduce the effect of cross-institution distribution differences. However, BERT relies on centralized corpora for model construction. It does not fully resolve semantic alignment or sensitive information separation in federated environments. Its performance is affected by differential privacy noise, heterogeneous text encoding, and inconsistencies in modeling objectives.

Compared with other methods, the proposed federated language framework achieves the best performance across all metrics. This demonstrates strong adaptability and robustness in cross-institution electronic medical record modeling. The model applies semantic factorization, cross-institution causal semantic alignment, and adaptive privacy budget control. These strategies reduce the effects of data heterogeneity, accumulated privacy noise, and semantic drift. The model retains stable and accurate medical text understanding while ensuring privacy protection. The overall results indicate that building a structured semantic space under privacy constraints is essential for improving the quality of federated electronic medical record processing.

This paper also presents a hyperparameter sensitivity experiment on the recall rate metric of the federal rounds cap, and the experimental results are shown in Figure 2.

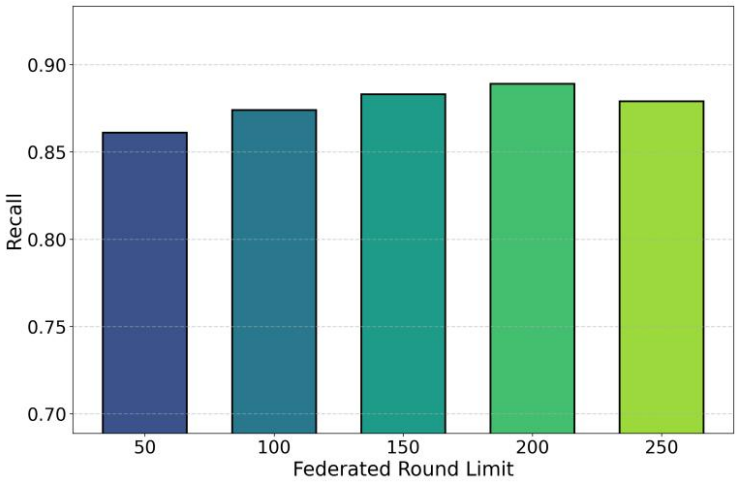

Figure 2. Hyperparameter sensitivity experiment of the federal rounds cap on the recall rate.

The experimental results show that the upper limit of federated rounds has a relatively steady influence on recall. When the number of rounds is low, the model has not fully aligned the cross-institutional semantic space. This leads to slightly lower recall. As communication rounds increase, the integration of local semantic factors into the global model becomes more complete. This allows the model to capture sparse symptom expressions and long-range semantic dependencies more effectively in cross-institution electronic medical records, which improves recall.

As the number of rounds continues to rise, recall becomes stable after reaching a certain point. This indicates that additional communication offers diminishing gains. This pattern is consistent with the characteristics of federated learning. In scenarios with strong cross-institution heterogeneity, early communication rounds reduce semantic distribution differences. Once the global semantic space stabilizes, more rounds focus mainly on local optimization and bring limited improvement. The results suggest that stable performance can be achieved with a reasonable number of rounds when semantic consistency and cross-institution alignment are ensured. It is unnecessary to pursue excessively high rounds.

The overall findings highlight the ability of the proposed federated language framework to balance communication efficiency and model performance. The combined effects of semantic factorization, cross-institution causal semantic alignment, and adaptive privacy budget control allow the model to maintain high recall with relatively few communication rounds. As the number of rounds increases moderately, the model strengthens its ability to capture implicit diagnostic cues in cross-institution electronic medical records. This demonstrates that the framework provides strong stability and convergence efficiency in real clinical environments.

This paper also presents an experiment on the environmental sensitivity of different privacy budget scheduling strategies to the AUC metric, and the experimental results are shown in Figure 3.

The experimental results show clear differences in how privacy budget scheduling strategies influence the AUC metric. This indicates that the dynamic allocation of privacy budgets directly affects the model's ability to capture semantic features of electronic medical records in federated settings. The fixed budget strategy produces the lowest performance. It cannot adjust noise levels based on gradient sensitivity or semantic complexity. As a result, noise may be excessive or insufficient during specific communication rounds. This weakens the model's semantic discrimination ability. The linear strategy performs better. It provides strong privacy protection in early stages and gradually increases model usability in later rounds.

The adaptive strategy further improves performance and achieves a slightly higher AUC than the linear strategy. It adjusts the budget according to gradient norms, semantic drift, or convergence status. This allows privacy noise to be applied more accurately during highly sensitive or uncertain stages. It reduces semantic information loss. The dynamic strategy achieves the highest AUC. This suggests that fine-grained and real-time budget adjustment can better match changes in semantic complexity across federated rounds. It helps the model maintain stable semantic representation under cross-institution conditions while minimizing performance degradation caused by privacy protection.

The hybrid strategy shows strong flexibility, but its AUC is slightly lower than that of the dynamic strategy. This indicates that the combination of multiple scheduling modes may lead to inconsistencies. The intensity of privacy noise may not fully align with changes in semantic features. Overall, the results show that stronger privacy budgets do not always lead to better outcomes in cross-institution electronic medical record modeling. Privacy budgets must be designed together with semantic factorization and structural alignment. A dynamic strategy that responds quickly to semantic gradient states achieves the best balance between privacy protection and semantic usability. It is therefore an essential component in building high-quality federated language models.

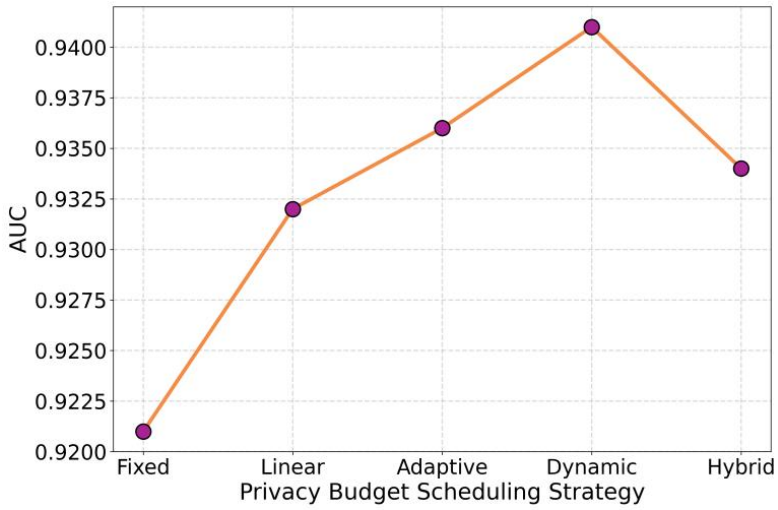

Figure 3. Environmental sensitivity experiment of different privacy budget scheduling strategies on the AUC metric

4. Conclusion

This study focuses on the core challenges of privacy protection, semantic consistency, and distribution heterogeneity in cross-institution collaboration using electronic medical records. It proposes a federated language modeling framework for medical text and introduces an adaptive privacy budget scheduling mechanism to achieve more robust cross-institution semantic integration. The method combines local text encoding, semantic factorization, cross-institution alignment, and dynamic privacy control. It builds a unified medical semantic space without sharing raw text. The framework overcomes the privacy compliance limitations of traditional centralized modeling. It also provides a new technical path for large-scale knowledge collaboration when data cannot flow across institutions.

The overall results show that the proposed method offers clear advantages in structured semantic representation, long-text understanding, and cross-institution modeling stability. By separating causal semantics from sensitive factors, the model reduces representation drift caused by variations in writing style and institutional documentation differences. The adaptive privacy budget mechanism aligns privacy

protection strength with training dynamics. It preserves task-relevant information while ensuring privacy. The design principles of this framework apply not only to electronic medical records but also to other privacy-sensitive professional text domains. They offer a scalable solution for high-sensitivity applications.

From an application perspective, this research supports cross-hospital collaboration, regional medical data-sharing networks, and the deployment of intelligent healthcare systems. As medical digitalization continues to grow, institutions have an increasing demand for secure and compliant text analytics. The proposed framework enables high-quality semantic understanding under federated conditions. It provides a reliable foundation for decision support, clinical prediction, and automated documentation tasks. The ideas of semantic decomposition and privacy budget scheduling can serve as technical references for future standardized privacy frameworks and promote the regulated use of privacy-preserving computation in medical text processing.

Future work will explore more fine-grained adaptive privacy mechanisms. The scheduling strategy will consider task difficulty, institutional heterogeneity, and model semantic uncertainty to achieve more targeted dynamic optimization. The federated language framework may also be extended to multimodal medical data. Integrating imaging data, structured measurements, and textual records could support a more comprehensive clinical knowledge space. With the rapid progress of large pretrained models in the medical domain, the proposed framework can also be combined with stronger foundation models. This will support the development of the next generation of privacy-preserving intelligent medical systems with improved reasoning, interpretability, and contextual adaptability. It will further enhance the usability and impact of medical artificial intelligence in real clinical environments.

References

[1] Cui J, Zhu H, Deng H, et al. FeARH: Federated machine learning with anonymous random hybridization on electronic medical records[J]. Journal of Biomedical Informatics, 2021, 117: 103735.

[2] Pan W, Xu Z, Rajendran S, et al. An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals[J]. Patterns, 2024, 5(1).

[3] Peng L, Luo G, Zhou S, et al. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction[J]. NPJ Digital Medicine, 2024, 7(1): 127.

[4] McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B. A., "Communication-efficient learning of deep networks from decentralized data", Proceedings of the Artificial Intelligence and Statistics, pp. 1273-1282, April 2017.

[5] Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review[J]. Computers in biology and medicine, 2023, 155: 106649.

[6] Chen C, Feng X, Li Y, et al. Integration of large language models and federated learning[J]. Patterns, 2024, 5(12).

[7] Wiest I C, Ferber D, Zhu J, et al. Privacy-preserving large language models for structured medical information retrieval[J]. NPJ Digital Medicine, 2024, 7(1): 257.

[8] Zheng J, Li J, Zhang Z, et al. Clinical Data based XGBoost Algorithm for infection risk prediction of patients with decompensated cirrhosis: A 10-year (2012 – 2021) Multicenter Retrospective Case-control study[J]. BMC gastroenterology, 2023, 23(1): 310.

[9] Si Y, Roberts K. Three-level hierarchical transformer networks for long-sequence and multiple clinical documents classification[J]. arXiv preprint arXiv:2104.08444, 2021.

[10] Prabhakar S K, Won D O. Medical text classification using hybrid deep learning models with multihead attention[J]. Computational intelligence and neuroscience, 2021, 2021(1): 9425655.

[11] Cheligeer C, Wu G, Lee S, et al. BERT-based neural network for inpatient fall detection from electronic medical records: retrospective cohort study[J]. JMIR medical informatics, 2024, 12: e48995.