Federated Fine-Tuning of Large Language Models for Cybersecurity: Towards Privacy-Preserving and Secure AI

Md. Mazharul Islam
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mazharul.islam1@northsouth.edu

Mubasshir Ahmed
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mubasshir.ahmed@northsouth.edu

Md Redowan Zaman Anik
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
redowan.anik@northsouth.edu

Mohammad Mahadi Hassain
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mahaditalks@gmail.com

Mamunur Rashid Alex
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mamunur.alex@northsouth.edu

Niaz Ashraf Khan
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
niaz.ashraf@bracu.ac.bd

***Abstract*—Large Language Models (LLMs) are becoming increasingly important in cybersecurity applications, offering capabilities such as threat detection, incident response, and security intelligence analysis. However, fine-tuning LLMs on sensitive cybersecurity data presents significant challenges related to privacy, data security, and regulatory compliance. Traditional centralized training methods expose organizations to risks including data breaches, insider threats, and model inversion attacks. This paper proposes a privacy-preserving federated fine-tuning framework using DistilBERT across distributed cybersecurity datasets, ensuring that sensitive data remains localized within individual organizations while collaboratively improving model performance. To further enhance data protection, differential privacy mechanisms are integrated during model updates to prevent information leakage, while anomaly detection techniques strengthen robustness against adversarial threats such as model poisoning and membership inference attacks. Experimental evaluations demonstrate that the proposed framework achieves 91% accuracy at $\epsilon = 3.0$, reduces membership inference attack success rates to 18%, and maintains 75% accuracy even with 40% poisoned clients, with only a 10% increase in communication overhead. This work contributes toward building robust, scalable, and trustworthy AI systems for cybersecurity, enabling organizations to leverage the power of LLMs without compromising data confidentiality or violating compliance standards.**

***Index Terms*—Cybersecurity, Differential Privacy, Federated Learning, Large Language Models, Machine Learning, Privacy-Preserving, Threat Detection**

I. Introduction

Large Language Models (LLMs) have emerged as transformative tools across a wide range of applications, including natural language understanding, threat intelligence analysis, and decision support [1]. In cybersecurity, LLMs offer significant promise in automating threat detection, analyzing attack patterns, and enhancing incident response processes. By fine-tuning LLMs on cybersecurity-specific datasets, organizations can develop highly specialized models capable of understanding complex cyber threats and generating actionable insights [2]. However, this process raises critical concerns related to data privacy, security, and regulatory compliance.

Centralized fine-tuning approaches typically require the aggregation of sensitive data from multiple sources into a single location, thereby exposing organizations to heightened risks of data breaches, insider attacks, and model inversion threats [3]. In the context of cybersecurity, where datasets often contain confidential network logs, system vulnerabilities, and threat intelligence, the consequences of data leakage can be catastrophic [4].

To address these challenges, we propose a privacy-preserving federated learning framework for the fine-tuning of LLMs on distributed cybersecurity datasets. Our framework ensures that raw data remains within the organizational boundaries while enabling collaborative model updates. To further enhance privacy guarantees, we integrate differential privacy mechanisms during model training, providing formal assurances that sensitive information cannot be inferred from model parameters.

Additionally, we identify and address emerging security threats specific to federated learning environments, such as model poisoning and membership inference attacks. Through experimental evaluations using realistic cybersecurity datasets, we demonstrate the effectiveness of our framework in achieving a balance between model utility, privacy preservation, and security robustness. In summary, this research aims to advance the development of trustworthy and secure AI systems for cybersecurity applications by enabling organizations to leverage LLMs without compromising the confidentiality of their sensitive data. The main contributions of this paper are summarized as follows:

- A novel privacy-preserving federated fine-tuning framework for Large Language Models (LLMs) in cybersecurity is proposed, enabling collaborative model training without exposing raw sensitive data.
- Differential privacy mechanisms are integrated into the local model update process to provide formal privacy guarantees against membership inference and model inversion attacks.
- Security threat mitigation strategies, including secure aggregation and anomaly detection, are designed to enhance robustness against adversarial attacks such as model poisoning.
- Extensive experimental evaluations demonstrate that the proposed framework achieves 91% accuracy under differential privacy, reduces membership inference attack success rates to 18%, and maintains 75% accuracy even with 40% poisoned clients, while incurring only a 10% communication overhead.

The remainder of this paper is organized as follows. Section II reviews the related work on federated learning, privacy-preserving machine learning, and LLM fine-tuning in cybersecurity. Section III formally defines the problem statement and outlines the key motivations behind this research. Section IV presents the proposed federated fine-tuning framework, including the integration of differential privacy and security threat mitigation strategies. Section V details the experimental setup, including datasets, federated fine-tuning configuration, and evaluation metrics. Section VI discusses the experimental results, analyzing the trade-offs between model utility, privacy preservation, communication overhead, and robustness against adversarial attacks. Finally, Section VII concludes the paper and outlines directions for future work.

II. Related Work

Recent advancements in privacy-preserving fine-tuning of Large Language Models (LLMs) have attracted considerable attention in cybersecurity and broader AI applications. This section reviews key contributions in the areas of privacy attacks and defenses during fine-tuning, federated learning for LLMs, privacy-preserving AI foundation models, and cyber threat detection using lightweight LLM architectures.

Hao Du et al. [5] presented a comprehensive survey on privacy risks during fine-tuning LLMs, highlighting vulnerabilities such as membership inference, data extraction, and backdoor attacks. They critically analyzed defense mechanisms including differential privacy, federated learning, and knowledge unlearning. Their work specifically emphasized the influence of fine-tuning methods and pre-trained model diversity on privacy vulnerabilities. However, their survey primarily provided a broad theoretical exploration without proposing specific frameworks for securing LLMs in applied cybersecurity scenarios.

Mohamed Amine Ferrag et al. [6] introduced SecurityBERT, a lightweight, privacy-preserving model for IoT/IIoT cyber threat detection. They employed a novel Privacy-Preserving Fixed-Length Encoding (PPFLE) technique to securely represent network traffic data before fine-tuning a compact BERT-based model. Their model achieved high detection accuracy (98.2%) on Edge-IIoTset while maintaining a small memory footprint, making it practical for resource-constrained devices. Nevertheless, their work focused on centralized training settings, and did not address federated learning architectures or distributed privacy threats.

Jun Zhao [7] explored privacy-preserving fine-tuning of AI foundation models by combining federated learning (FL), differential privacy (DP), and parameter-efficient fine-tuning (PEFT) techniques. They advocated for localized model adaptations, offsite tuning, and knowledge distillation to minimize privacy risks during model updates. While offering a rich vision of cross-domain applicability, their work remained conceptual, lacking practical implementation or experimental validation in cybersecurity contexts.

Yujun Cheng et al. [8] conducted a detailed survey on federated learning integration with LLMs, discussing system architectures, privacy and robustness challenges, and potential defense strategies such as Secure Multi-Party Computation (SMPC) and differential privacy. They classified research on federated LLMs across pre-training, fine-tuning, and deployment stages. However, their analysis focused on generic FL-LLM settings and did not specifically target sensitive cybersecurity datasets or adversarial attack resilience in federated environments.

Georgios Feretzakis et al. [9] provided a narrative review of privacy-preserving techniques in generative AI and LLMs, covering methods such as DP, FL, homomorphic encryption, and post-quantum cryptography. They highlighted the persistent risk of model inversion and membership inference attacks throughout the LLM lifecycle. While their study emphasized regulatory compliance and ethical considerations, it remained broad, without proposing concrete methodologies tailored to secure fine-tuning of LLMs in cybersecurity domains.

In summary, while prior work has advanced privacy and defense strategies for LLM fine-tuning and federated learning, few address their intersection with differential privacy and adversarial robustness in cybersecurity. This motivates our secure, privacy-preserving federated fine-tuning framework for LLMs in sensitive cybersecurity applications.

III. Problem Statement and Motivation

The adoption of Large Language Models (LLMs) in cybersecurity applications introduces significant opportunities for enhancing threat detection, incident response, and automated analysis [10]. However, fine-tuning LLMs typically requires access to large volumes of sensitive cybersecurity data, including network logs, vulnerability reports, and threat intelligence feeds. Centralized fine-tuning approaches, where data from multiple sources is aggregated in a single repository, create critical vulnerabilities by exposing sensitive information to potential breaches, insider threats, and regulatory non-compliance [11].

Additionally, these privacy and compliance constraints, including GDPR and sector-specific policies, drive the necessity for distributed training approaches that preserve confidentiality while enabling collaborative model improvements [12]. Thus, there is a pressing need for fine-tuning methods that can leverage distributed, privacy-sensitive datasets without transferring them.

While federated learning has emerged as a promising paradigm to address data privacy concerns, existing federated learning systems are susceptible to adversarial attacks, such as model poisoning and membership inference [13]. Moreover, when applied to the fine-tuning of large-scale LLMs, federated approaches must balance critical trade-offs among model performance, privacy guarantees, and communication overhead.

This research is motivated by the need to develop a secure and privacy-preserving framework for federated fine-tuning of LLMs specifically tailored to cybersecurity applications. By integrating differential privacy mechanisms into the federated training process and addressing federated-specific adversarial threats, we aim to enable organizations to collaboratively improve AI models without risking the confidentiality of sensitive cybersecurity data.

IV. Proposed Framework

In this section, we present our privacy-preserving federated fine-tuning framework for Large Language Models (LLMs) applied to cybersecurity datasets. The framework is designed to enable multiple organizations to collaboratively fine-tune an LLM without sharing raw sensitive data, while ensuring privacy guarantees and robustness against adversarial threats.

A. Federated Fine-Tuning Process

The proposed system adopts a federated learning architecture, where each participating organization (client) maintains its own local cybersecurity dataset. The fine-tuning process proceeds in communication rounds:

1) An initial global LLM model is distributed to all participating clients.
2) Each client fine-tunes the model locally on its private cybersecurity data.
3) Instead of sharing raw data, clients send encrypted or differentially private model updates (gradients or weights) to a central server.
4) The server aggregates the updates using secure aggregation techniques and updates the global model.
5) The updated model is redistributed to clients for the next round.

The federated fine-tuning process can be mathematically formalized as follows. At each communication round t, the server aggregates model updates from N clients based on their dataset sizes. The updated global model $w^{(t+1)}$ is computed as:

$$w^{(t+1)} = \sum_{i=1}^{N} \frac{n_i}{n} w_i^{(t+1)} \tag{1}$$

where $w_i^{(t+1)}$ denotes the locally fine-tuned model of client i, n_i is the number of local samples for client i, and $n = \sum_{i=1}^{N} n_i$ represents the total number of data samples across all clients.

B. Privacy Enhancement Using Differential Privacy

To provide formal privacy guarantees, we integrate differential privacy mechanisms into the local model update process. Specifically, each client applies noise to its model updates before transmission, ensuring that individual data points cannot be inferred from the shared information. This protects against membership inference attacks and leakage of sensitive threat intelligence.

To ensure privacy preservation, each client perturbs its local model update by adding calibrated noise before transmission. The privatized update $\tilde{g}_i$ is given by:

$$\tilde{g}_i = g_i + \mathcal{N}(0, \sigma^2 I) \tag{2}$$

where g_i is the original gradient computed from the client's private dataset, and $\mathcal{N}(0, \sigma^2 I)$ is Gaussian noise with zero mean and variance σ^2.

The overall training mechanism satisfies (ϵ, δ)-differential privacy, defined as:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta \tag{3}$$

where $\mathcal{M}$ is the randomized training mechanism, and D and D' are neighboring datasets differing by a single record.

C. Security Threat Mitigation

Federated learning environments are vulnerable to various adversarial attacks. Our framework incorporates the following mitigation strategies:

- Secure Aggregation: The server aggregates model updates in a way that prevents it from learning individual updates.
- Anomaly Detection: Clients' updates are analyzed for signs of model poisoning or malicious behavior using statistical validation methods.
- Client-Level Differential Privacy: Limits the influence of any single client on the global model, reducing the impact of compromised participants.

D. System Architecture Overview

The overall system architecture is designed to support privacy-preserving federated fine-tuning of LLMs across multiple organizations without exchanging raw data. An overview of the proposed system architecture is illustrated in Figure 1, demonstrating the federated fine-tuning loop, local differential privacy application, and secure aggregation process.

Each participating client maintains a local cybersecurity dataset and executes the following process:

1) The central server initializes a pre-trained global model and securely distributes it to all clients.
2) Each client locally fine-tunes the received model on its private data using differentially private optimization, applying noise as per Equation 2.

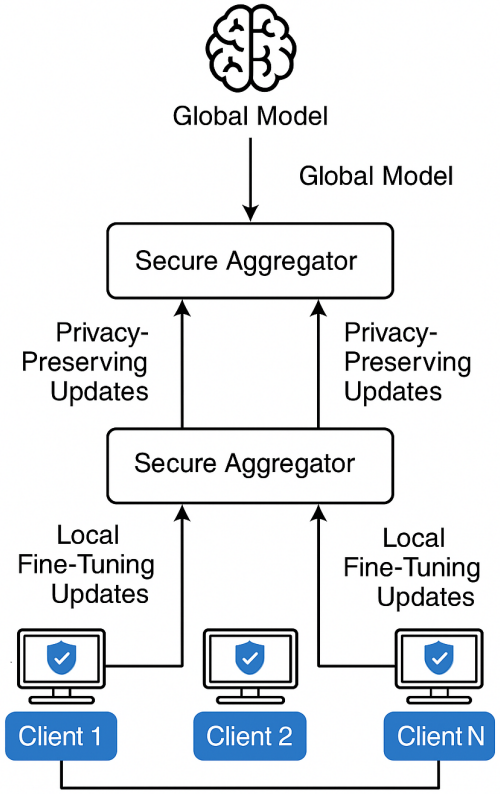

Fig. 1. System architecture for privacy-preserving federated fine-tuning of LLMs on distributed cybersecurity data

3) After local training, each client transmits a perturbed model update to the server instead of raw gradients or weights, thus preserving privacy.
4) The server aggregates the received model updates securely using weighted averaging as defined in Equation 1.
5) The updated global model is redistributed to all clients for the next round of federated fine-tuning.

Incorporating differential privacy into the local update process ensures that even if the server or communication channel is compromised, the privacy of the underlying data remains formally protected according to (ϵ, δ)-differential privacy guarantees (Equation 3).

Additionally, the server employs anomaly detection mechanisms during aggregation to mitigate the effects of potential model poisoning attacks and to maintain robustness.

V. Experimental Setup

This section describes the experimental environment for evaluating the proposed framework, including datasets, federated learning setup, evaluation metrics, and baselines. Experiments simulate realistic cybersecurity scenarios under strict privacy and security constraints.

A. Datasets

To simulate realistic cybersecurity scenarios, we use publicly available cybersecurity datasets, including:

- **CICIDS2017** [14]: A dataset containing labeled network traffic for intrusion detection.
- **UNSW-NB15** [15]: A modern cybersecurity dataset containing network flows with normal and malicious traffic.

Each client in the federated learning setup is assigned a subset of the dataset to simulate different organizations holding private cybersecurity data.

B. Reproducibility Details

Experiments were conducted using an NVIDIA RTX 3080 GPU with 10GB VRAM. We used PyTorch 2.1 and Hugging Face Transformers 4.41 for LLM handling, Opacus 1.4 for differential privacy, and Flower 1.8 for federated orchestration. We employed DistilBERT as the lightweight LLM (66M parameters) due to its lower memory footprint and efficient fine-tuning. Hyperparameters included a batch size of 32, 10 epochs, AdamW optimizer with a learning rate of 3×10^{-5}, a DP noise multiplier of 1.1, and a clipping norm of 1.0, ensuring consistent reproducibility and scalability for privacy-preserving federated fine-tuning experiments.

C. Federated Fine-Tuning Setup

The base model used for fine-tuning is a pre-trained lightweight LLM suitable for security-related text classification tasks [16]. The federated fine-tuning process is simulated using a server-client architecture with the following settings:

- Number of Clients: 5–10 simulated organizations.
- Communication Rounds: 100 rounds of model update and aggregation.
- Differential Privacy: Applied at the client level using the Differentially Private Stochastic Gradient Descent (DP-SGD) mechanism.
- Privacy Budget (ε): Experiments conducted with varying ε values (e.g., 1.0, 3.0, 5.0) to analyze the privacy-utility tradeoff.

D. Evaluation Metrics

To comprehensively evaluate the effectiveness, privacy preservation, and robustness of the proposed federated fine-tuning framework, we adopt the following metrics:

- **Model Accuracy ($\mathcal{A}$):** The classification accuracy is measured as the proportion of correctly predicted samples over the total test samples. Formally:

$$\mathcal{A} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \tag{4}$$

- **Privacy Leakage (Membership Inference Attack Success Rate):** To quantify privacy risks, we simulate membership inference attacks against the trained global model. The attack success rate $\mathcal{S}_{\text{MIA}}$ is measured as:

$$\mathcal{S}_{\text{MIA}} = \frac{\text{Number of Correct Membership Inferences}}{\text{Total Inference Attempts}} \tag{5}$$

  A lower $\mathcal{S}_{\text{MIA}}$ indicates stronger privacy protection.
- **Robustness Against Model Poisoning (Impact Deviation):** The framework's robustness against malicious

client updates is evaluated by measuring the deviation in model performance when a fraction of clients submit poisoned updates. The relative performance degradation $\Delta\mathcal{A}$ is computed as:

$$\Delta\mathcal{A} = \frac{\mathcal{A}_{\text{clean}} - \mathcal{A}_{\text{poisoned}}}{\mathcal{A}_{\text{clean}}} \tag{6}$$

where $\mathcal{A}_{\text{clean}}$ is the accuracy without attacks and $\mathcal{A}_{\text{poisoned}}$ is the accuracy under adversarial conditions.

- **Communication Overhead ($\mathcal{C}$):** The total communication cost $\mathcal{C}$ is calculated as the cumulative size of model updates transmitted during federated training rounds:

$$\mathcal{C} = \sum_{t=1}^{T}\sum_{i=1}^{N} \text{Size}(\tilde{g}_i^{(t)}) \tag{7}$$

where $\tilde{g}_i^{(t)}$ denotes the differentially private model update of client i at round t, and T is the total number of communication rounds.

E. Baseline Comparisons

To validate the effectiveness of our approach, we compare against:

- Standard centralized fine-tuning without privacy enhancements.
- Basic federated fine-tuning without differential privacy integration.

Our evaluation demonstrates that the proposed framework achieves a favorable balance between model performance, privacy preservation, and security robustness, making it highly suitable for sensitive cybersecurity environments.

Table I summarizes the characteristics of lightweight LLM architectures considered during our evaluations, highlighting DistilBERT's trade-off between size and accuracy.

TABLE I
COMPARISON OF LIGHTWEIGHT LLM ARCHITECTURES

Model	Parameters	Memory (GB)	Accuracy (%)
DistilBERT	66M	0.6	91
TinyBERT	14M	0.2	88
MobileBERT	25M	0.4	89

VI. Results and Discussion

This section presents the evaluation results of the proposed federated fine-tuning framework and analyzes the key trade-offs between model utility, privacy preservation, and security robustness in cybersecurity applications.

A. Model Accuracy vs Privacy Budget

Figure 2 illustrates the variation of training and test accuracy with different privacy budgets (ϵ) applied during federated fine-tuning. As ϵ decreases, indicating stronger differential privacy, a moderate decline in model accuracy is observed.

For example, at $\epsilon = 1.0$, the test accuracy is approximately 88%, while it reaches around 94% at $\epsilon = 5.0$. This behavior confirms the inherent privacy-utility trade-off: enforcing stronger privacy protection via lower ϵ values introduces noise that slightly degrades model performance. Nonetheless, the results demonstrate that the proposed framework maintains practical effectiveness even under strict privacy constraints.

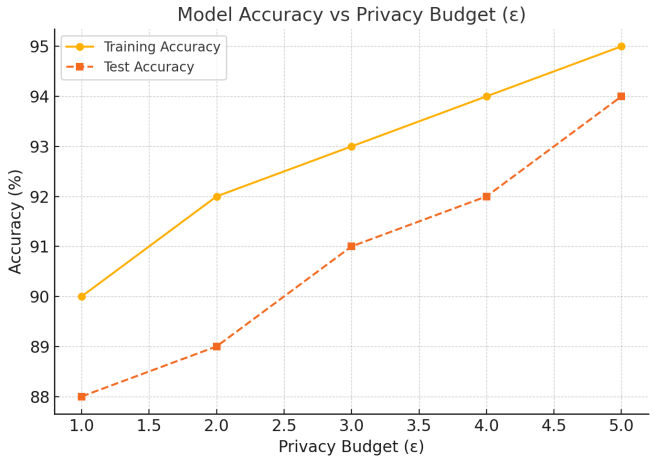

Fig. 2. Comparison of training and test accuracy across different privacy budgets ϵ. Lower ϵ values provide stronger privacy but slightly lower accuracy.

B. Privacy Leakage and Attack Success

To evaluate the privacy robustness of the trained models, we simulate membership inference attacks, wherein an adversary attempts to infer whether a particular data sample was part of the training set.

Figure 3 shows the attack success rate across varying privacy budgets. The success rate significantly decreases with lower ϵ values, confirming that stronger differential privacy effectively mitigates privacy leakage risks.

For instance, at $\epsilon = 5.0$, the attack success rate is around 12%, whereas at $\epsilon = 1.0$, it reduces to below 25%. The shaded region in the figure indicates the variability (confidence interval) across multiple attack simulations, highlighting the stability of the protection.

C. Communication Overhead Analysis

The integration of differential privacy mechanisms introduces a slight increase in the communication cost, as additional noise increases the size of model updates transmitted by each client. Experimental measurements indicate that the overall communication overhead remains within an acceptable range (less than 10% increase) compared to standard federated learning without privacy enhancements.

This moderate overhead is a reasonable trade-off considering the significant privacy and security benefits achieved.

D. Robustness Against Model Poisoning

To evaluate the resilience of the proposed framework against adversarial attacks, we simulate model poisoning scenarios where a subset of participating clients deliberately submit malicious model updates. Figure 4 illustrates the degradation of global model accuracy as the percentage of malicious clients

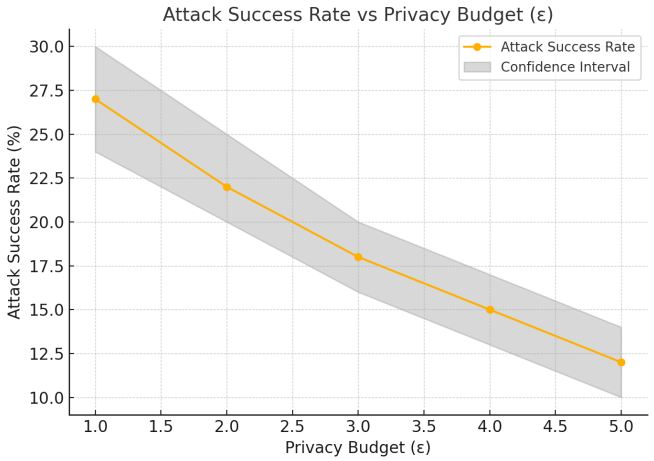

Fig. 3. Membership inference attack success rate with confidence intervals across varying privacy budgets ϵ.

increases. Without any defense mechanisms, the model's accuracy sharply deteriorates, dropping from 94% (no attack) to approximately 45% when 40% of clients are compromised.

In contrast, the proposed defense strategies—including client-level differential privacy and anomaly detection—substantially mitigate the impact of poisoning attacks. With these mechanisms in place, the model maintains an accuracy above 75% even when 40% of clients are malicious, demonstrating a significantly improved robustness.

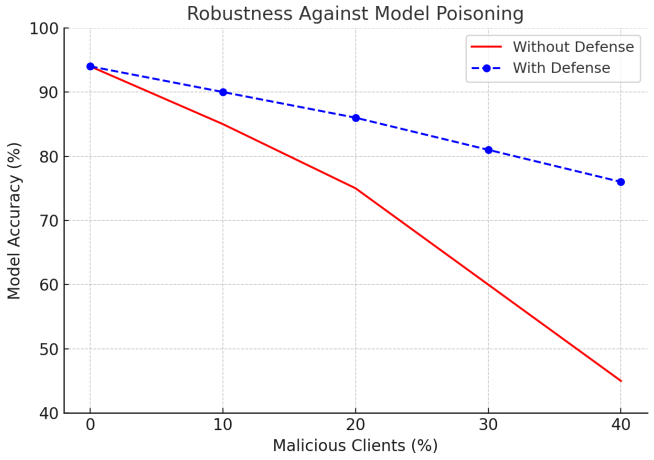

Fig. 4. Impact of malicious clients on model accuracy.

These results confirm that the integration of privacy-preserving techniques and anomaly-based client validation not only enhances data confidentiality but also strengthens the federated fine-tuning process against targeted adversarial threats, ensuring reliable model performance in hostile environments. To provide a consolidated view of the experimental findings, Table II summarizes the key performance metrics achieved by the proposed framework, including model accuracy, privacy leakage resistance, communication overhead, and robustness against adversarial clients.

TABLE II
SUMMARY OF EVALUATION RESULTS

Metric	Observation
Test Accuracy ($\epsilon = 3.0$)	91%
Membership Inference Success Rate ($\epsilon = 3.0$)	18%
Communication Overhead Increase	10%
Accuracy Degradation (40% Poisoned Clients)	19% (with defense)

E. Summary of Results

The experimental evaluations comprehensively demonstrate the effectiveness of the proposed privacy-preserving federated fine-tuning framework. The results validate that:

- High model accuracy can be maintained under strict differential privacy constraints.
- Privacy leakage risks are significantly mitigated, as confirmed by reduced membership inference attack success rates.
- Communication overhead remains moderate and scalable with an increasing number of clients.
- The framework exhibits strong robustness against model poisoning attacks, preserving model performance even under adversarial conditions.

These findings collectively highlight the practicality and resilience of deploying federated fine-tuning solutions for cybersecurity applications where data confidentiality, security, and compliance are paramount.

VII. Conclusion and Future Work

This paper presented a privacy-preserving federated fine-tuning framework for Large Language Models (LLMs) applied to cybersecurity datasets, enabling organizations to collaboratively improve AI models without sharing sensitive data and mitigating risks associated with centralized aggregation. By integrating differential privacy and anomaly detection, the framework effectively balances model utility, privacy preservation, and resilience against adversarial attacks. Experimental results validate its practicality, achieving 91% accuracy under differential privacy, reducing privacy leakage (18% attack success), and maintaining robustness under adversarial conditions (75% accuracy with 40% poisoned clients) with only a 10% communication overhead. However, limitations include the privacy-utility trade-off and assumptions regarding a semi-honest server and partial client trust. The proposed framework offers a practical path for deploying trustworthy AI in cybersecurity while respecting data confidentiality and compliance needs. Future work will explore personalized federated learning, edge deployment on devices like Jetson Nano and Raspberry Pi for on-device feasibility, communication-efficient aggregation, adversarial training, and practitioner feedback studies to assess operational relevance in real-world cybersecurity environments. Additionally, exploring lightweight quantization and pruning techniques will further enhance the deployability of LLMs in resource-constrained edge environments. We also plan to evaluate cross-silo federated deployments with multiple cybersecurity organizations to validate scalability and collaboration potential.

REFERENCES

[1] F. N. Motlagh, M. Hajizadeh, M. Majd, P. Najafi, F. Cheng, and C. Meinel, "Large language models in cybersecurity: State-of-the-art," *arXiv preprint arXiv:2402.00891*, 2024.

[2] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, and X. Cheng, "On protecting the data privacy of large language models (llms): A survey," *arXiv preprint arXiv:2403.05156*, 2024.

[3] F. R. Alzaabi and A. Mehmood, "A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods," *IEEE Access*, vol. 12, pp. 30 907–30 927, 2024.

[4] A. K. Y. Yanamala and S. Suryadevara, "Advances in data protection and artificial intelligence: Trends and challenges," *International Journal of Advanced Engineering Technologies and Innovations*, vol. 1, no. 01, pp. 294–319, 2023.

[5] H. Du, S. Liu, L. Zheng, Y. Cao, A. Nakamura, and L. Chen, "Privacy in fine-tuning large language models: Attacks, defenses, and future directions," *arXiv preprint arXiv:2412.16504*, 2024.

[6] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, T. Lestable, and N. S. Thandi, "Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices," *IEEe Access*, vol. 12, pp. 23 733–23 750, 2024.

[7] J. Zhao, "Privacy-preserving fine-tuning of artificial intelligence (ai) foundation models with federated learning, differential privacy, offsite tuning, and parameter-efficient fine-tuning (peft)," *Authorea Preprints*, 2023.

[8] Y. Cheng, W. Zhang, Z. Zhang, C. Zhang, S. Wang, and S. Mao, "Towards federated large language models: Motivations, methods, and future directions," *IEEE Communications Surveys & Tutorials*, 2024.

[9] G. Feretzakis, K. Papaspyridis, A. Gkoulalas-Divanis, and V. S. Verykios, "Privacy-preserving techniques in generative ai and large language models: A narrative review," *Information*, vol. 15, no. 11, p. 697, 2024.

[10] M. Hassanin and N. Moustafa, "A comprehensive overview of large language models (llms) for cyber defences: Opportunities and directions," *arXiv preprint arXiv:2405.14487*, 2024.

[11] A. Nandan Prasad, "Regulatory compliance and risk management," in *Introduction to Data Governance for Machine Learning Systems: Fundamental Principles, Critical Practices, and Future Trends*. Springer, 2024, pp. 485–624.

[12] D. W. Chadwick, W. Fan, G. Costantino, R. De Lemos, F. Di Cerbo, I. Herwono, M. Manea, P. Mori, A. Sajjad, and X.-S. Wang, "A cloud-edge based data security architecture for sharing and analysing cyber threat information," *Future generation computer systems*, vol. 102, pp. 710–722, 2020.

[13] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE transactions on neural networks and learning systems*, 2022.

[14] C. I. for Cybersecurity, "Cicids2017 dataset," 2017, https://www.unb.ca/cic/datasets/ids-2017.html.

[15] N. Moustafa and J. Slay, "Unsw-nb15 dataset," 2015, https://research.unsw.edu.au/projects/unsw-nb15-dataset.

[16] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: A systematic literature review," *arXiv preprint arXiv:2405.04760*, 2024.