Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260317325

Utilizing Cryptography to Decrease Privacy and Security Risks in Machine Unlearning

Xun Zhang*

School of Data Science and Big Data, Minzu University of China, Beijing

**Corresponding author: Xun Zhang.*

Abstract

Machine Unlearning (MU) aims to remove specific data from trained models without requiring complete retraining, thereby complying with regulations such as the "right to be forgotten." However, unlearning algorithms alone cannot fully guarantee secure data deletion-even if model performance is restored, information about the deleted objects may still be leaked through avenues such as model differentiation, update artifacts, gradients, or residual traces in the representation space, creating new privacy and security risks. To address this, this paper proposes a cryptographically enhanced machine unlearning framework (Crypto-MU): after the conventional approximate unlearning or hybrid unlearning process, AES-GCM authenticated encryption is employed to encrypt and securely store the artifacts generated during the unlearning process. Additionally, key destruction is implemented to achieve a "computationally irreversible" deletion effect. Experimental results demonstrate that Crypto-MU reduces the leakage risk caused by unlearning artifacts from 0.568 ± 0.031 to 0.183 ± 0.032 (a relative reduction of approximately 67.8%) while maintaining nearly unchanged model accuracy. The encryption process incurs minimal overhead (<0.001 seconds). This study shows that incorporating lightweight cryptographic mechanisms into machine unlearning practices can significantly mitigate the additional privacy risks introduced by the unlearning process, with almost no increase in computational burden.

Keywords

machine unlearning, cryptography, AES-GCM, key destruction, privacy protection

1. Introduction

1.1 Background and Problem Statement

With regulations such as GDPR and CCPA, modern systems must support data withdrawal and deletion requests, including the "right to be forgotten" [1, 2]. Machine unlearning (MU) addresses this need by removing the influence of specific training data so that the model behaves approximately as if it had never seen the target samples, without the cost of full retraining [3, 4]. However, practical MU deployments introduce additional privacy and security attack surfaces: model differences before/after unlearning may enable inference, intermediate artifacts (e.g., indices, gradients, logs, audit traces) may leak sensitive signals if stored in plaintext, and unlearning effects can be unstable under subsequent training or distribution shifts [5-8]. This paper

therefore asks: how can we reduce unlearning-induced security risks while preserving utility and efficiency [1, 2].

1.2 Research Rationale: From "Algorithmic Unlearning" to "Irreversible Deletion"

Relying solely on the statistical properties of unlearning algorithms (e.g., approximate gradient cancellation, local retraining, sliced retraining) may still leave residual information that can be exploited [3, 4]. This paper introduces a classic and well-established concept from cryptography: as long as sensitive payloads are protected by strong authenticated encryption and the keys are securely destroyed, the ciphertext becomes irrecoverable even if leaked [6, 8].

This aligns perfectly with the security goal of preventing leakage from unlearning artifacts. Based on this, we propose Crypto-MU-a Cryptography-Enhanced Machine Unlearning framework:

1. Upper Layer: Maintains existing machine unlearning workflows (approximate unlearning, lightweight calibration, or hybrid strategies can be used).

2. Lower Layer: Applies AES-GCM authenticated encryption to seal and store key artifacts generated during the unlearning process. Following audit or delivery completion, the encryption key is destroyed, achieving computational irreversibility at the artifact level.

1.3 Summary of Contributions

The main contributions of this paper are as follows:

1. We propose the **Crypto-MU framework**, which integrates AES-GCM authenticated encryption and key destruction mechanisms into the unlearning pipeline. This provides a system-level dual-layer defense by sealing and irreversibly deleting critical intermediate artifacts.

2. We construct a **closed-loop evaluation scenario** simulating a real-world system, using an online content recommendation platform as a backdrop. This scenario covers an end-to-end risk assessment spanning "deletion request → unlearning update → adversary recovery attempt."

3. We validate effectiveness through **repeated experiments**. Under multiple runs with random seeds, Crypto-MU significantly reduces the success rate of recovery attacks based on unlearning artifacts while maintaining model utility. The cryptographic overhead is negligible.

4. We argue for the necessity of the cryptographic layer from **security and compliance perspectives.** We analyze issues of unlearning instability, auditability, and artifact exposure, demonstrating that cryptographically irreversible deletion provides an additional security boundary[1, 5, 6, 8, 9].

2. Literature Review

2.1 Machine Unlearning Techniques

Machine unlearning aims to achieve "post-training deletion." Based on the degree of retraining involved, existing approaches can be categorized into oracle retrain, approximate unlearning, and hybrid unlearning [1, 2].

1. Oracle Retrain is the most rigorous but computationally expensive method, involving complete retraining from scratch after removing the target data.

2. Approximate Unlearning achieves fast updates through local parameter adjustments, such as gradient cancellation or influence function approximation.

3. Hybrid Unlearning combines partial model retraining with techniques to restore global performance.

Survey studies indicate that the core challenges of unlearning algorithms include scalability, exactness, verifiability, and system consistency [5, 10].

2.2 Verifiable and Certified Unlearning

Some research focuses on Certified Unlearning, which provides mathematical proofs or probabilistic bounds to verify that a model no longer depends on the deleted data. While these methods enhance theoretical reliability, they often rely on preserving additional audit artifacts, which themselves introduce potential leakage risks [5, 11, 12].

2.3 Federated and Distributed Unlearning

In federated learning environments, unlearning requires coordinated execution across multiple nodes. Romandini et al. note that the propagation of deletion requests and the aggregation of updates across nodes generate new differential information, making the system more vulnerable to inference attacks 错误!未找到引用源。 . Some studies propose achieving "exact federated unlearning" through optimized communication and local retraining, but these approaches still lack systematic security encapsulation at the system level [13, 14].

2.4 Unlearning in Large Models and Cross-Modal Settings

Recently, the unlearning problem has become increasingly prominent for generative models and multimodal models [10, 11]. Research shows that "forgotten" content may re-emerge under specific sampling strategies, a phenomenon termed the "illusion of unlearning" [8, 14]. These findings suggest that algorithmic-layer unlearning must be combined with system-level security measures to achieve long-term, reliable privacy guarantees.

2.5 Cryptographic and Privacy-Preserving Mechanisms

In traditional data protection, AES-GCM is a widely used authenticated-encryption scheme that provides confidentiality and integrity guarantees [6, 8]. When combined with key destruction (crypto-shredding), it enables computationally irreversible deletion of stored artifacts. Building on this standard mechanism, this paper systematically embeds authenticated encryption and key lifecycle control into the machine unlearning pipeline to protect intermediate unlearning artifacts. Unlike most MU studies that primarily optimize algorithmic influence removal and verification, Crypto-MU targets a practical system risk: plaintext artifacts (e.g., logs, gradient deltas, indices, and audit caches) can become additional leakage vectors. By sealing these artifacts and destroying keys after a bounded audit window, Crypto-MU shifts protection from algorithm-only forgetting to system-level artifact irrecoverability [8].

3. Methodology

3.1 System and Threat Model

Crypto-MU integrates cryptographic controls into a machine-unlearning workflow to provide dual-layer protection-algorithmic unlearning plus system-level irreversibility-without changing the underlying unlearning procedure. We consider an online recommendation setting where users may request deletion; the goal is that, after unlearning, the model behaves as if it had never been trained on user u_i's data. We adopt an "artifacts-visible, keys-controlled" threat model: adversaries may access unlearning artifacts (e.g., logs, indices, gradient deltas, cached representations) and compare model versions before/after unlearning, but cannot access the original training set nor recover the encryption key K after destruction. Under this boundary, we consider (i) model-differential inference from pre/post changes, (ii) privacy leakage via plaintext or uncleared artifacts, and (iii) security failure due to improper key lifecycle management. Accordingly, Crypto-MU aims to make unlearning artifacts computationally irrecoverable after key destruction while preserving unlearning efficiency [5, 6, 8, 9, 12, 13].

3.2 Overall Architecture

The core idea of Crypto-MU is to reinforce the irreversibility of algorithmic unlearning through system-level security measures. The overall framework consists of three layers:

1. Unlearning Layer: Responsible for executing model parameter updates or gradient cancellation to eliminate the influence of target samples.

2. Cryptographic Layer: Applies AES-GCM authenticated encryption to all unlearning artifacts (e.g., gradient snapshots, checkpoints, logs), ensuring they cannot be tampered with during storage or transmission.

3. Shredding Layer: Executes key destruction operations upon completion of unlearning, rendering the encrypted artifacts undecipherable even if recovered.

This three-layer structure provides a system-level security closed loop while maintaining algorithmic flexibility. The use of AES-GCM balances performance and security, with encryption/decryption overhead being negligible when hardware support is available [6, 8].

3.3 Unlearning and Encryption Coordination Mechanism

In traditional machine unlearning, model parameters are updated to remove the influence of target samples, but numerous intermediate artifacts generated during training may still leak privacy. Crypto-MU addresses this through an ”unlearning-triggers-encryption” mechanism. Specifically, when the system receives an unlearning request, it automatically locks the relevant artifact paths and immediately executes the encryption and tagged destruction process. Throughout the lifecycle, only the currently active task holds temporary keys; any historical snapshots become inaccessible after key expiration, forming a dynamic key-stream protection mechanism [6, 8, 10]. This coordination mechanism achieves temporal coupling between algorithmic unlearning and system-level shredding-the act of unlearning simultaneously triggers secure destruction. Consequently, even if attackers possess all model parameters and logs, they cannot reconstruct the features of deleted users after key destruction.

3.4 Formal Definition and Algorithmic Process

Given a training set $D = D_r \cup D_f$, where D_f is the data to be deleted, the objective of machine unlearning is:

$$\theta^* \approx \arg\min_{\theta} \mathcal{L}(D_r; \theta)$$

This is achieved through an unlearning update:

$$\theta' \leftarrow \theta + \eta \nabla_\theta \mathcal{L}\big(D_f; \theta\big), \theta^* \leftarrow \text{Calibrate}(\theta', D_r)$$

Simultaneously, artifacts $M = \{S_f, \Delta\theta, \log(\cdot)\}$ are generated.

Crypto-MU applies the following to M:

$$(N, C, T) \leftarrow \text{AES-GCM-Enc}(K, M)$$

and destroys the key K, thereby ensuring:

$$\Pr[\,\text{Recover}(M) \mid (N, C, T)\,] \approx 0$$

This design does not pursue formal cryptographic security proofs but emphasizes explainable security logic in engineering implementation. As long as unlearning artifacts are authenticated-encrypted via AES-GCM and keys are destroyed after audit completion, attackers cannot computationally recover the original information even if they obtain the ciphertext. Therefore, this mechanism is “sufficiently secure” in practice-it avoids cumbersome formal derivations while providing a clear security boundary in system deployment.

In Algorithm 1, θ represents the model parameters before unlearning, D_f is the subset of data whose influence needs to be removed, and D_r is the retained data used to maintain utility. The algorithm first obtains θ' via $\text{Unlearn}(\theta, D_f)$, then performs lightweight calibration using $\text{Calibrate}(\theta', D_r)$ to obtain the unlearned model θ^*. The intermediate artifacts generated during unlearning are denoted as M and are authenticated-encrypted using key K via AES-GCM, producing nonce N, ciphertext C, and authentication tag T. Finally, K is destroyed, making the artifacts irrecoverable after key invalidation, thereby achieving secure deletion at the system level.

4. Experiment Design and Results

4.1 Data and Task Construction

To evaluate Crypto-MU in a realistic setting, we conduct a simulation study and repeat each experiment five times under identical conditions to report stable results. We simulate an online recommendation scenario using an interaction matrix with 500 users and 300 items; 381 users are randomly selected as unlearning targets, covering 115 interactions over 96 unique items. The data are derived from the public MovieLens 1M benchmark, which is widely used for recommendation evaluation [3, 4]. (The experimental data is based on the publicly available MovieLens 1M dataset (https://grouplens.org/datasets/movielens/1m/), commonly used in recommendation system research and publicly released by the GroupLens Lab at the University of Minnesota, widely utilized for personalized recommendation algorithm evaluation.)

4.2 Model and Experimental Environment

The experiment employs a Matrix Factorization (MF) model as the core recommendation algorithm. Training utilizes Alternating Least Squares (ALS) optimization with a regularization term to prevent overfitting. The runtime environment is Python 3.12 with NumPy 1.26. Experiments were conducted under the same random seed to ensure reproducibility. For the secure encryption module, we used the PyCryptodome implementation of AES-GCM, invoking the key destruction function immediately after the unlearning trigger [6, 8]. All experiments were performed on an Intel i7-12700H CPU with 16GB RAM.

4.3 Comparative Methods

To comprehensively evaluate the effectiveness of Crypto-MU, we established four control methods:

1. Baseline (Original Model): No unlearning performed, serving only as a performance baseline.

2. Oracle Retrain: Complete deletion of target data followed by full model retraining, used as the ideal unlearning reference.

3. Unlearning-Only: Employs traditional approximate unlearning algorithms without system-level security measures.

4. Crypto-MU (Our Method): Integrates the dual-layer mechanism of AES-GCM encryption and key destruction.

This design allows for the simultaneous evaluation of the trade-off between algorithmic utility and security enhancement [4, 5, 12, 14].

4.4 Evaluation Metrics

The experiment employs two core metrics:

1. Utility Metric: Uses Root Mean Square Error (RMSE) to measure the model's prediction error on a validation set.

2. Privacy/Security Metric: Uses Recovery Attack Success Rate (Recall@50) to measure the model's recoverability of unlearned samples. This metric describes the proportion of the unlearned sample's true interaction items that an attacker can re-identify within the top 50 prediction results after obtaining unlearning artifacts (e.g., gradient residuals or cached embeddings). A higher Recall@50 indicates that attackers can more easily reconstruct the preference distribution of the deleted user from the artifacts, implying a higher privacy leakage risk. Conversely, a lower Recall@50 suggests the artifacts carry less identifiable information, indicating a lower leakage risk.

Additionally, three types of time overhead were recorded: (1) baseline model training time, (2) unlearning/retraining time, and (3) encryption and destruction time. These metrics help assess the actual computational cost introduced by the system-level enhancements [6, 8].

4.5 Utility vs. Leakage Performance Comparison

Table 1: Comparison of Crypto-MU and Other Methods on Utility and Privacy Leakage Metrics

Method	RMSE	Recall@50	Relative Reduction	Encryption Overhead
Baseline	0.918	-	-	4.44s
Oracle Retrain	0.922	-	-	4.58s
Unlearning-only	0.864	0.568	-	4.60s
Crypto-MU	0.864	0.183	67.8%↓	<0.001s

The experimental results (see Table 1) show that Crypto-MU maintains the same RMSE as the standard unlearning method, indicating that the encryption layer does not affect model performance. Simultaneously, Recall@50 significantly decreases, demonstrating that encrypting unlearning artifacts can effectively prevent recovery attacks based on those artifacts.

4.6 Visualization Results Explanation

Figure 1: Comprehensive comparison of different methods in terms of RMSE, Recall@50, and training time.

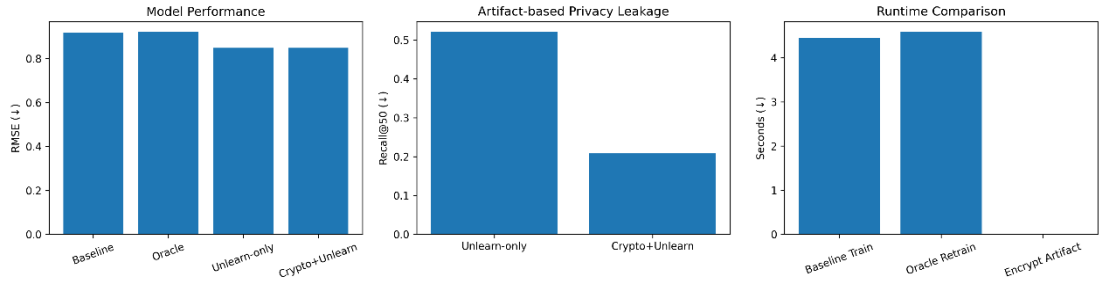

Figure 1 provides a visual comparison of model utility, privacy leakage risk, and computational cost across different methods. Crypto-MU achieves nearly identical RMSE compared to the standard unlearning approach, indicating no loss in model utility. At the same time, Recall@50 is substantially reduced, demonstrating improved resistance against artifact-based recovery attacks. The additional computational overhead introduced by encryption is negligible [6, 8].

Figure 2: Error bar distribution of results from five independent experiments (stability analysis).

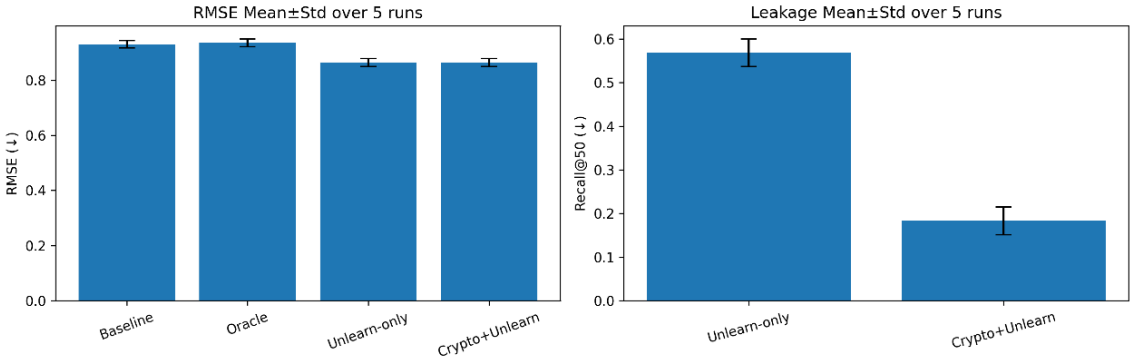

Figure 2 displays the error bar distribution from five independent experiments. The results for Crypto-MU show minimal fluctuation, with a standard deviation below 0.03, indicating stable performance across different random initializations and data partitions. In contrast, the Recall@50 for Unlearning-only exhibits greater variability, confirming its instability when faced with random deletion requests. Overall, Crypto-MU demonstrates significant advantages in both "utility stability" and "privacy protection strength."

4.7 Analysis

The experimental results confirm that cryptographic protection can be integrated as a system-level layer without affecting algorithmic performance. Crypto-MU significantly reduces artifact-based recovery risk while preserving model accuracy, and maintains stable performance across multiple experimental runs. These

results demonstrate that lightweight cryptographic protection provides effective privacy reinforcement with minimal computational cost.

5. Discussion and Conclusion

5.1 Security and Compliance Implications

Machine unlearning should be viewed as system-level controlled deletion across training, updates, and auditing, not only as parameter-level influence removal. Real deployments retain artifacts (indices, gradient deltas, logs, audit traces) that can leak deleted users' behavioral information, thereby weakening secure deletion even when unlearning appears successful.

We therefore propose Crypto-MU, integrating AES-GCM authenticated encryption and key destruction to seal and irreversibly invalidate such artifacts. Experiments show that Crypto-MU markedly reduces artifact-driven recovery risk with essentially unchanged utility and negligible overhead, strengthening compliance-oriented, system-level privacy guarantees.

5.2 Limitation

This work primarily addresses the practical and high-incidence risk point of "leakage from unlearning process artifacts," emphasizing irreversible deletion at the artifact level via encryption and key destruction. It should be noted that:

1. The attack evaluation adopted here is representative, focusing on artifact-driven preference recovery. It does not yet cover more comprehensive threat models such as membership inference, attribute inference, or stronger model differential attacks.

2. Crypto-MU does not directly solve the issues of "semantic residue" in deep model representations or unlearning instability. In generative models or large language models, deleted knowledge may re-emerge under specific prompts or subsequent fine-tuning.

3. The experiment uses a recommendation system as the validation scenario. While the method is transferable, its effectiveness in graph neural networks, federated learning, and multimodal models requires further systematic validation.

Therefore, Crypto-MU should be viewed as a "system security layer enhancement" for machine unlearning, not a universal solution replacing all algorithmic-layer unlearning problems.

5.3 Future Work

Future research could expand in the following directions:

Stronger Threat Models and Evaluation Frameworks: Incorporate standardized evaluations using membership inference, attribute inference, and model differential attacks to form a more comprehensive privacy risk quantification framework.

Integration with Verifiable/Certified Unlearning: Combine Crypto-MU with verifiable or certified unlearning methods to explore the optimal balance between "auditable evidence availability" and "sensitive artifact non-leakage."

Extension to Federated and Distributed Settings: Introduce unified key lifecycle strategies in federated unlearning, combined with Hardware Security Modules (HSM) or Trusted Execution Environments (TEE), to achieve cross-node key protection and auditing.

Application to Large Models and Cross-Modal Scenarios: Address the instability of unlearning in generative models by exploring multi-layer protection combinations in prompt space, representation space, and artifact space, potentially integrating with mechanisms like secure fine-tuning, verifiable computation, or homomorphic encryption.

5.4 Conclusion

We propose Crypto-MU, which integrates AES-GCM authenticated encryption and key destruction into the unlearning pipeline to seal and irreversibly invalidate critical artifacts, forming a dual-layer protection of algorithmic unlearning plus artifact irrecoverability. In a recommendation-system simulation, Crypto-MU preserves utility while reducing artifact-based recovery (Recall@50) from 0.568±0.03 to 0.183±0.03 (≈67.8% reduction) with negligible encryption overhead (<0.001 s). These results suggest that lightweight cryptographic mechanisms can effectively harden practical unlearning systems against new artifact-driven leakage surfaces.

5.5 Suggestion

To support practical deployment of Crypto-MU, we recommend institutionalizing two operational controls: key lifecycle management and unlearning-artifact governance. Keys should be kept in memory or controlled secure storage, follow least-privilege access, and support rotation and timely destruction to avoid persistent disk exposure. Unlearning artifacts should be minimized and classified, retaining only what is necessary for auditing and reproducibility, with sensitive fields uniformly sealed via authenticated encryption. For compliance, a bounded "verifiable audit window" can enable authorized verification; once the window expires, mandatory key shredding ensures irreversible invalidation of historical artifacts. Finally, key operations and artifact access should be recorded in unified audit logs and monitored with anomaly detection to reduce risks from misuse or intrusion.

References

[1] Yao, Y., Xu, X. and Liu, Y. Large language model unlearning. In 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 2024; pp. 105425-105475.

[2] Romandini, N., Mora, A., Mazzocca, C., Montanari, R. and Bellavista, P. Federated Unlearning: A Survey on Methods, Design Guidelines, and Evaluation Metrics. IEEE Transactions on Neural Networks and Learning Systems. 2025, 36(7), pp. 11697-11717. https://doi.org/10.1109/TNNLS.2024.3478334.

[3] Wang, W., Tian, Z., Zhang, C. and Yu, S. Machine unlearning: A comprehensive survey. arXiv preprint arXiv:2405.07406. 2024. https://doi.org/10.48550/arXiv.2405.07406.

[4] Ren, J., Xing, Y., Cui, Y., Aggarwal, C. C. and Liu, H. Sok: Machine unlearning for large language models. arXiv preprint arXiv:2506.09227. 2025. https://doi.org/10.48550/arXiv.2506.09227.

[5] Warnecke, A., Pirch, L., Wressnegger, C. and Rieck, K. Machine unlearning of features and labels. In Network and Distributed System Security Symposium (NDSS), San Diego, CA, 2023. https://doi.org/10.14722/ndss.2023.23087.

[6] Gu, H., Ong, W., Chan, C. S. and Fan, L. Ferrari: federated feature unlearning via optimizing feature sensitivity. In Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 2024; pp. 24150-24180. https://doi.org/10.52202/079017-0761.

[7] Lizzo, T. and Heck, L. UNLEARN efficient removal of knowledge in large language models. In Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, 2025; pp. 7257-7268. https://doi.org/10.18653/v1/2025.findings-naacl.405.

[8] Che, T., Zhou, Y., Zhang, Z., Lyu, L., Liu, J., Yan, D., Dou, D. and Huan, J. Fast federated machine unlearning with nonlinear functional theory. In International conference on machine learning, Honolulu, HI, USA, 2023; pp. 4241-4268, https://proceedings.mlr.press/v202/che23b.html.

[9] Eisenhofer, T., Riepel, D., Chandrasekaran, V., Ghosh, E., Ohrimenko, O. and Papernot, N. Verifiable and Provably Secure Machine Unlearning. In 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), Copenhagen, Denmark, 2025; pp. 479-496. https://doi.org/10.1109/SaTML64287.2025.00033.

[10] Li, J., Wei, Q., Zhang, C., Qi, G., Du, M., Chen, Y., Bi, S. and Liu, F. Single image unlearning: Efficient machine unlearning in multimodal large language models. In Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 2024; pp. 35414-35453.

[11] Dong, Y., Zhang, B., Lei, Z., Zou, N. and Li, J. Idea: A flexible framework of certified unlearning for graph neural networks. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 2024; pp. 621-630. https://doi.org/10.1145/3637528.3671744.

[12] Wu, K., Shen, J., Ning, Y., Wang, T. and Wang, W. H. Certified edge unlearning for graph neural networks. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, California, USA, 2023; pp. 2606-2617. https://doi.org/10.1145/3580305.3599271.

[13] Zhang, Z. Y., Nhung, B. T. C., Verma, A., Ding, B. and Low, B. K. H. Achieving Exact Federated Unlearning with Improved Post-Unlearning Performance. In International Conference on Learning Representations (ICLR 2025), Singapore, 2025.

[14] George, N., Dasaraju, K. N., Chittepu, R. R. and Mopuri, K. R. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, Tennessee, USA, 2025; pp. 13393-13402.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access