Incentivizing Truthful Machine Unlearning via Hierarchical Auditing

Shaolong Guo , **Yuntao Wang*** , **Zhou Su*** and **Tom H. Luan**
School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, China
shaolong.guo@stu.xjtu.edu.cn, {yuntao.wang, tom.luan}@xjtu.edu.cn, zhousu@ieee.org

Abstract

Machine unlearning has become a critical capability for AI services to comply with evolving privacy regulations. A key yet underexplored challenge is how to *verify* whether a profit-driven AI server has faithfully performed unlearning. Existing verification approaches either incur prohibitive costs or provide insufficient deterrence, failing to balance audit cost and enforcement effectiveness. To bridge this gap, we propose UAG, a game-theoretic unlearning auditing framework that incentivizes truthful unlearning via strategic deterrence rather than exhaustive verification. We design a hierarchical auditing mechanism that combines low-cost screening with selectively triggered high-precision verification, and models the server–auditor interaction as a three-stage dynamic Bayesian game. By characterizing the equilibrium strategy, we derive optimal audit and penalty policies that incentivize honest unlearning. Theoretical analysis and experiments show that UAG maintains reliable detection while achieving a favorable cost–deterrence tradeoff. Notably, UAG attains a server honesty rate of approximately 95% while screening only about 50% of unlearning requests, showing its practicality for trustworthy black-box unlearning services.

1 Introduction

With the rapid commercialization of artificial intelligence (AI), particularly the surge in large language models, machine learning as a service (MLaaS) has become a dominant paradigm for deploying high-value models. In domains such as healthcare, finance, and personalized AI services, MLaaS platforms routinely train AI models on massive volumes of sensitive user data. As regulatory frameworks such as the GDPR and the CCPA increasingly enforce the *right to be forgotten*, *machine unlearning* (MU) has emerged as a critical capability for MLaaS [Guo *et al.*, 2025]. It mandates service providers to permanently remove the influence of specific data samples from deployed models upon request.

Although various MU algorithms have been proposed [Lee *et al.*, 2025; Lyu *et al.*, 2025; He *et al.*, 2025; Yang *et al.*, 2025; Li *et al.*, 2025], most existing studies primarily focus on improving computational efficiency and minimizing accuracy degradation. However, algorithmic unlearning does not imply verifiable unlearning. In large-scale MLaaS systems, faithful unlearning often incurs substantial economic costs (e.g., tens to hundreds of GPU hours) and can cause non-negligible performance loss. This creates strong incentives for profit-driven service providers to claim compliance (i.e., unlearning has been executed) while strategically avoiding costly unlearning operations. Hence, ensuring *verifiability* of unlearning has become a fundamental prerequisite for establishing trust among users, auditors, and MLaaS providers.

Recent studies have explored verifiable unlearning mechanisms to assess whether requested data has been properly removed, typically by detecting residual information associated with forgotten samples. Existing approaches can be broadly categorized into three classes. *Trigger-based methods* [Sommer *et al.*, 2022; Guo *et al.*, 2024; Han *et al.*, 2025] inject carefully crafted samples during training and verify unlearning by probing the model's responses to these triggers. However, such methods are fundamentally limited in practice: users generally cannot anticipate which samples will require unlearning after deployment, making it infeasible to pre-embed triggers for targeted data during training [Wang *et al.*, 2025b]. Moreover, trigger injection may degrade model utility and compromise the privacy of unrelated data contributors [Zhang *et al.*, 2024; Xu *et al.*, 2026].

To better align with practical deployments, recent works have shifted towards *post-hoc verification* approaches compatible with MLaaS. Among them, *membership inference-based methods* [Chen *et al.*, 2021; Golatkar *et al.*, 2021; Jia *et al.*, 2023; Kurmanji *et al.*, 2023] assess whether forgotten samples remain distinguishable from unseen data, whereas *reconstruction-based methods* [Xu *et al.*, 2026; Wang *et al.*, 2025b] quantify behavioral differences before and after unlearning. However, these methods face a fundamental *cost–accuracy–security dilemma*. High-precision verification techniques can improve detection accuracy but incur prohibitive computational and operational costs if applied frequently. In contrast, lightweight checks are inexpensive yet offer limited deterrence and can be strategically bypassed by a profit-driven service provider. Besides, unlearning verifica-

*Zhou Su and Yuntao Wang are the corresponding authors.

tion is inherently strategic, as the server's compliance behavior adapts to the auditor's verification frequency, detection accuracy, and penalty structure. It remains an open challenge to design an efficient unlearning auditing mechanism that balances verification cost, detection effectiveness, and robustness to strategic manipulation.

To bridge this gap, this paper proposes the first practical machine Unlearning Audit Game (UAG) for MLaaS systems, which explicitly models unlearning verification as a strategic interaction between a profit-driven server and a resource-constrained auditor. By theoretically characterizing equilibrium behavior, UAG provides principled guidance for resolving the *cost–accuracy–security* dilemma in real-world unlearning scenarios. Specifically, UAG consists of the following two tightly coupled components.

1) Hierarchical Unlearning Audit Game. We model the strategic interaction between the server and the auditor as a three-stage dynamic Bayesian game under incomplete information. The core idea is to integrate scalable, low-cost screening mechanisms (e.g., membership inference attack (MIA)-based test [Jia *et al.*, 2023]) as a first-line audit, with high-precision but expensive verification (e.g., reconstruction-based test [Wang *et al.*, 2025b]) that is selectively triggered upon detecting suspicious signals. This strategic audit structure transforms costly verification from a constant operational burden into a credible probabilistic threat. By embedding heterogeneous audit mechanisms within a unified game-theoretic framework, UAG achieves effective strategic deterrence while preserving audit scalability in practical MLaaS deployments.

2) Equilibrium Analysis and Optimal Strategy Design. Within the UAG framework, we develop an efficient analytical approach to jointly derive equilibrium strategies, including the auditor's optimal audit frequency and the server's optimal compliance probability. Through rigorous equilibrium analysis, we reveal how the auditor should calibrate audit intensity and penalty levels to balance deterrence and operational cost. Specifically, the equilibrium analysis identifies two practically relevant operating regions. In the *audit-light equilibrium*, sufficiently strong penalties and accurate screening make dishonest unlearning unattractive, allowing the auditor to rely primarily on low-cost screening and trigger high-precision verification only upon alert signals, thereby maintaining low auditing overhead without sacrificing security. In contrast, when penalties are insufficient or screening reliability is too low to deter strategic non-compliance, the system shifts to an *audit-intensive equilibrium*, in which the auditor should increase the frequency of high-precision audits to incentivize honest unlearning.

We conduct extensive experiments to compare UAG with state-of-the-art schemes, including MIA [Jia *et al.*, 2023] and TAPE [Wang *et al.*, 2025b]. The results show that UAG achieves a favorable cost–deterrence trade-off, maintaining a server honesty rate of around 0.95 with a basic audit probability of around 0.5. UAG also provides higher audit reliability and substantially lower overhead than existing schemes. Overall, our study offers practical insights into designing incentive-compatible unlearning auditing mechanisms with strong security guarantees and controlled operational costs.

2 Related Works

Verifiable MU Approaches. Existing verifiable MU approaches aim to provide evidence that the influence of specific data has been effectively removed from a trained model. Hardware-assisted approaches leverage trusted execution environments to certify unlearning operations [Weng *et al.*, 2024], offering strong correctness guarantees while relying on specialized hardware and facing scalability limitations.

In parallel, a dominant line of work verifies unlearning through attack-based detection. For instance, backdoor-based schemes embed verification triggers during training and assess unlearning by probing model responses [Sommer *et al.*, 2022; Guo *et al.*, 2024; Hu *et al.*, 2022; Han *et al.*, 2025], whereas inference-based approaches evaluate whether the model's predictive confidence on a target sample decays to a level indistinguishable from unseen data [Chen *et al.*, 2021; Golatkar *et al.*, 2021; Jia *et al.*, 2023; Kurmanji *et al.*, 2023]. More recently, reconstruction-based methods quantify unlearning by examining behavioral discrepancies between pre- and post-unlearning models [Xu *et al.*, 2026; Wang *et al.*, 2025b]. However, these approaches either require intrusive modifications to the training pipeline or offer limited deterrence against strategic non-compliance. In contrast, our framework integrates low-cost screening with high-precision verification, achieving effective deterrence without disrupting the model training process.

Mechanism Design for MU. Game-theoretic modeling has recently been introduced to MU to address both internal algorithmic trade-offs and external incentive design. From an algorithmic perspective, prior studies formulate unlearning as a strategic interaction among competing objectives. For example, [Wu and Harandi, 2025] applies Nash bargaining to balance forgetting and model utility, while [Liu *et al.*, 2025] characterizes the trade-off between unlearning effectiveness and privacy leakage to limit additional risks. From an economic perspective, existing studies design incentives to regulate unlearning-related behaviors of servers and users. Representative works employ Stackelberg games to price data retention and reduce unlearning costs [Cui and Cheung, 2025], use contract theory to encourage data contribution through unlearning guarantees [Wang *et al.*, 2025a] or to balance personalized privacy and model utility [Pan *et al.*, 2026], and apply evolutionary games to motivate resource sharing for collaborative retraining [Lin *et al.*, 2024]. Most existing studies implicitly assume faithful execution of unlearning by the server and overlook strategic deviations in practice. In contrast, our work models the server as a strategic player and introduces a Bayesian auditing mechanism to deter non-compliance under incomplete information.

3 Model

As depicted in Fig. 1, we consider an MU setting with three entities: an ML server S, a set of data providers $\mathcal{I} = \{1, \ldots, I\}$, and a third-party auditor A. Each provider i contributes a local dataset $\mathcal{D}_i$ with size $d_i = |\mathcal{D}_i|$. The server aggregates all datasets $\mathcal{D} = \bigcup_{i \in \mathcal{I}} \mathcal{D}_i$ to train a global model Ψ and offers prediction services [Wang *et al.*, 2021].

Nota.	Description
x	Server's probability of honest unlearning execution
$\mathcal{V}_b, \mathcal{V}_e$	Basic audit and extended audit
α	Auditor's probability of initiating basic audit
s	Screening signal obtained from basic audit
$\mu(s)$	Posterior belief of server's cheating given signal s
γ_0, γ_1	Extended audit probabilities given safe and alert signals
η_k, ϵ_k	True positive and false positive rates of audit $k \in \{b, e\}$

Table 1: Summary of key notations used.

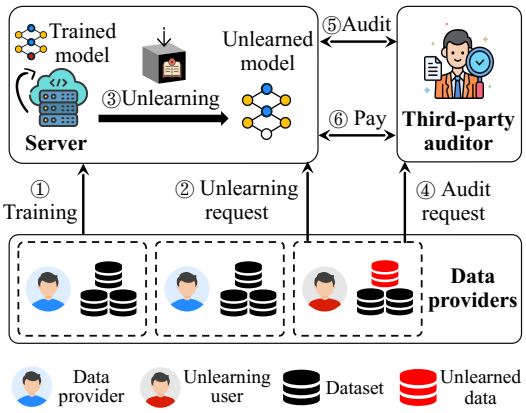

Figure 1: Illustration of machine unlearning with an auditor.

Unlearning Objective. For privacy or regulatory reasons, each provider $i \in \mathcal{I}$ may request the removal of a subset $\mathcal{D}_i^u \subseteq \mathcal{D}_i$ of size d_i^u. Upon receiving unlearning requests, the server is expected to release an unlearned model Ψ_u^* that eliminates the influence of the target data while preserving the utility of the remaining data. Ideally, the optimal unlearned model Ψ_u^* can be obtained by minimizing the empirical risk on the remaining data [Wang *et al.*, 2024]:

$$\Psi_u^* = \arg\min_{\Psi} \sum_{i \in \mathcal{I}} \frac{d_i - d_i^u}{D - D^u} \mathcal{L}(\Psi; \mathcal{D}_i \setminus \mathcal{D}_i^u), \qquad (1)$$

where $\mathcal{L}(\cdot)$ denotes a task-specific loss function, $D = \sum_i d_i$ is the total training data size, and $D^u = \sum_i d_i^u$ represents the total volume requested to be removed.

Server's Action. In practice, obtaining Ψ_u^* may require costly retraining and degrade model performance, creating incentives for the server to deviate from honest unlearning. We model the server as a strategic agent choosing an action $a \in \{a_h, a_c\}$, where a_h denotes faithful unlearning and a_c denotes cheating by skipping the unlearning process.

Auditor's Action. The auditor verifies unlearning compliance (i.e., whether the server performs truthful unlearning) under limited resources. High-precision verification ensures reliability but is economically infeasible at scale, whereas lightweight screening is inexpensive yet vulnerable to evasion [Sommer *et al.*, 2022]. To balance audit accuracy and cost, a hierarchical audit mechanism is devised as follows.

Definition 1 (Hierarchical Audit Mechanism)**.** *The mechanism $\mathcal{M} = \langle \mathcal{V}_b, \mathcal{V}_e \rangle$ combines low-cost screening (basic audit) with selectively triggered high-precision verification (extended audit):*

i) Basic audit. *With probability $\alpha \in [0, 1]$, the auditor conducts a lightweight screening $\mathcal{V}_b$, which yields a binary signal $s \in \mathcal{S} = \{s_0, s_1\}$, corresponding to negative (safe) and positive (alert), respectively. The screening is imperfect and characterized by true positive rate η_b (i.e., $\Pr(s_1|a_c)$) and false positive rate ϵ_b (i.e., $\Pr(s_1|a_h)$). We can instantiate $\mathcal{V}_b$ via MIA, which provides fast and non-intrusive statistical evidence of whether the removed data still influences the model. While efficient, such tests are inherently noisy and cannot provide definitive guarantees [Sommer* et al.*, 2022].*

ii) Extended audit. *Conditioned on signal s, the auditor may escalate to $\mathcal{V}_e$ following a signal-dependent strategy $\Gamma = \{\gamma_0, \gamma_1\}$, where $\gamma_j = \Pr(\mathcal{V}_e|s_j)$ for $j \in \{0, 1\}$. Compared with basic screening, extended audits are more expensive but provide near-definitive outcomes, characterized by a higher true positive rate $\eta_e > \eta_b$ and a lower false positive rate $\epsilon_e < \epsilon_b$. The audit $\mathcal{V}_e$ outputs a verdict $v \in \{v_p, v_f\}$, corresponding to* pass *or* fail*. The extended audit can be instantiated using a high-precision verification method, such as TAPE [Wang* et al.*, 2025b], which reconstructs residual information from model posterior differences to assess whether the requested data has been effectively removed.*

Consequently, the auditing process terminates in one of four outcomes $\mathcal{V} = \{v_\emptyset, v_t, v_p, v_f\}$: no audit, termination after screening, passing the extended audit, or failing the extended audit with a fine.

4 Unlearning Audit Game

4.1 Motivation and Game Overview

The interaction between the server and the auditor is inherently strategic under information asymmetry. The server holds private knowledge of its unlearning execution status and may bypass costly unlearning when the perceived audit risk is low. In contrast, the auditor aims to enforce compliance while controlling verification costs.

Game Process. We model this strategic interaction as a multi-stage dynamic Bayesian game. As illustrated in Fig. 2, the game proceeds in three stages. First, the server chooses its execution strategy, deciding whether to perform honest unlearning or to cheat. Second, the auditor decides whether to perform a basic audit, which yields a preliminary verification signal. Third, after observing the signal, the auditor updates its belief and decides whether to escalate to an extended audit. By analyzing this game, we characterize the optimal responses of both parties and derive the auditor's optimal audit policy, including the audit frequency and escalation rule, to effectively deter server cheating.

Definition 2 (Unlearning Audit Game (UAG))**.** *The strategic interaction is formally modeled as a dynamic Bayesian game $\mathcal{G} = \langle \mathcal{N}, \Phi, \mathcal{U} \rangle$, characterized by:*

- ***Players*** *($\mathcal{N}$): The player set is $\mathcal{N} = \{S, A\}$, representing the server and auditor, respectively.*
- ***Strategy spaces*** *(Φ): i) The server's strategy ϕ_S is parameterized by the honesty probability x, defined as*

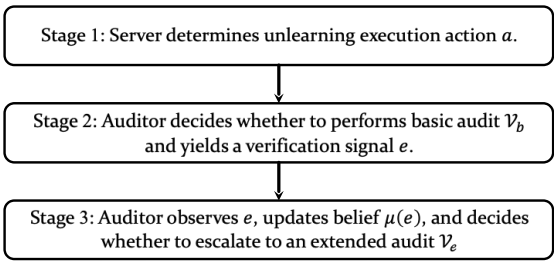

Figure 2: Three-stage dynamic Bayesian game.

$\Pr(a = a_h)$. *ii) The auditor's strategy is a tuple* $\phi_A = (\alpha, \Gamma)$*, comprising the basic screening probability* α *and extended audit probabilities* $\Gamma = \{\gamma_0, \gamma_1\}$.

- ***Payoffs*** $(\mathcal{U})$***:*** *The utilities of the server and the auditor are denoted by* U_S *and* U_A*, respectively, and their expected utilities are denoted by* $\mathbb{U}_S$ *and* $\mathbb{U}_A$.

4.2 Server's Utility

The server aims to maximize the residual utility of the deployed model while minimizing unlearning cost and potential penalties. Its utility is defined as

$$U_S(a, v) = V_{\text{res}}(a) - C_{\text{exe}}(a) - \Phi_{\text{pen}}(v). \tag{2}$$

i) Residual model utility. Honest unlearning typically induces a non-linear degradation in model accuracy [Cui and Cheung, 2025; Jiao *et al.*, 2021]. To capture this, we model the residual utility V_{res} using an exponential decay form:

$$V_{\text{res}}(a) = \begin{cases} \delta V_{\max}, & \text{if } a = a_c, \\ \delta[V_{\max} - (A_1 z^{A_2 \sum_i d_i^u} - A_3)], & \text{otherwise,} \end{cases} \tag{3}$$

where δ converts model performance into economic utility, and A_1, A_2, A_3, z are curve-fitting constants obtained from empirical findings [Cui and Cheung, 2025].

ii) Unlearning execution cost. Honest execution incurs computational costs (e.g., retraining time), typically linear in the remaining data size [Bourtoule *et al.*, 2021; Chen *et al.*, 2021; Warnecke *et al.*, 2023]:

$$C_{\text{exe}}(a) = \begin{cases} 0, & \text{if } a = a_c, \\ c_0 \left(\sum_{i \in \mathcal{I}} d_i - \sum_{i \in \mathcal{I}} d_i^u \right), & \text{otherwise,} \end{cases}$$

where c_0 denotes unit computational cost [Wang *et al.*, 2022].

iii) Conviction penalty. If convicted, the server pays a fixed fine, expressed as

$$\Phi_{\text{pen}}(v) = \mathbb{I}(v = v_f) V_{\text{pen}}. \tag{4}$$

4.3 Auditor's Utility

The auditor aims to deter cheating while controlling audit costs and decision risks (including both false negatives and false positives). Its utility is given by

$$U_A(a, v) = \Phi_{\text{pen}}(v) - C_{\text{ver}}(v) - L_{\text{err}}(a, v). \tag{5}$$

i) Conviction revenue. Upon issuing the conviction verdict v_f, the auditor receives penalty revenue Φ_{pen} (cf. Eq. (4)).

ii) Hierarchical audit cost. The operational cost depends on audit depth. Let c_b and c_e denote the costs of basic and extended verification ($c_e > c_b$), respectively, we have

$$C_{\text{ver}}(v) = c_b \mathbb{I}(v \neq v_\emptyset) + c_e \mathbb{I}(v \in \{v_p, v_f\}).$$

iii) Error loss. The auditor incurs a loss for incorrect decisions. We quantify the risks of failing to penalize a cheating server incurs λ_{fn}, while wrongfully convicting an honest server incurs λ_{fp}, and is given by:

$$L_{\text{err}}(a, v) = \begin{cases} \lambda_{\text{fn}}, & \text{if } a = a_c \wedge v \neq v_f, \\ \lambda_{\text{fp}}, & \text{if } a = a_h \wedge v = v_f, \\ 0, & \text{otherwise.} \end{cases}$$

We assume $\lambda_{\text{fp}} > V_{\text{pen}}$ to discourage wrongful convictions.

5 Equilibrium Analysis

The analysis proceeds as follows: we first characterize the auditor's belief update (Sec. 5.1) and optimal extended audit policy in Stage III (Sec. 5.2), which partition the game into three regions. Then, we derive the equilibrium strategies for both players within each region (Secs. 5.3–5.5) and present an equilibrium strategy derivation algorithm in Sec. 5.6.

5.1 Posterior Belief Update

In Stage III, the auditor observes a realization of the screening signal $s \in \mathcal{S}$ and updates its belief about the server's compliance according to Bayes' rule. Let $\mu(s) \triangleq \Pr(a_c|s)$ denote the posterior belief that the server has cheated. Specifically, when the screening returns an alert signal s_1, the auditor derives the posterior suspicion as

$$\begin{aligned} \mu(s_1) &= \frac{\Pr(s_1|a_c)\Pr(a_c)}{\Pr(s_1|a_c)\Pr(a_c) + \Pr(s_1|a_h)\Pr(a_h)} \\ &= \frac{\eta_b(1-x)}{\eta_b(1-x) + \epsilon_b x}. \end{aligned} \tag{6}$$

Similarly, upon observing s_0, the posterior belief is

$$\mu(s_0) = \frac{(1-\eta_b)(1-x)}{(1-\eta_b)(1-x) + (1-\epsilon_b)x}. \tag{7}$$

Given that the basic audit is informative (i.e., $\eta_b > \epsilon_b$) [Sommer *et al.*, 2022], an alert signal induces a higher posterior suspicion than a safe signal. Consequently, the monotonicity property $\mu(s_1) > \mu(s_0)$ holds for any x.

5.2 Optimal Extended Audit Policy

In Stage III, after observing s, the auditor applies the escalation strategy $\Gamma = \{\gamma_0, \gamma_1\}$ to decide whether to trigger the extended audit. We first derive the critical belief threshold that governs this escalation decision.

Lemma 1 (Investigation Threshold)**.** *There exists a critical belief* τ^* *such that the auditor escalates to an extended audit* ($\gamma^* = 1$) *if the posterior suspicion* $\mu(s) > \tau^*$*, terminates the audit if* $\mu(s) < \tau^*$*, and is indifferent when* $\mu(s) = \tau^*$*. This threshold is given by*

$$\tau^* = \frac{c_e + \epsilon_e(\lambda_{\text{fp}} - V_{\text{pen}})}{\eta_e(V_{\text{pen}} + \lambda_{\text{fn}}) + \epsilon_e(\lambda_{\text{fp}} - V_{\text{pen}})}.$$

Region	Honesty range	Risk level	Optimal policy $\Gamma^* = (\gamma_0^*, \gamma_1^*)$
I	$x \in [0, \hat{x}_{\text{safe}})$	High	$(1, 1)$
II	$x \in (\hat{x}_{\text{safe}}, \hat{x}_{\text{alert}})$	Medium	$(0, 1)$
III	$x \in (\hat{x}_{\text{alert}}, 1]$	Low	$(0, 0)$

Table 2: Optimal extended audit policy across different regions.

Remark: All proofs are provided in the online appendix [Guo *et al.*, 2026a]; see Appendix 1 for the proof of Lemma 1. Intuitively, τ^* is the break-even belief at which the auditor's expected gain from detecting a violation balances the cost of the extended audit and the risk of wrongful conviction.

While τ^* is defined in the *belief space*, the server controls the *strategy space* via the honesty parameter x. Since $\mu(s)$ is strictly decreasing in x, we can translate τ^* into thresholds over x by inverting the Bayesian update functions in Eqs. (6)–(7). The resulting thresholds divide the server's strategy space into regions with different optimal escalation decisions.

Definition 3 (Strategy-Space Boundaries)**.** *The server's strategy space $x \in [0, 1]$ is divided by two critical honesty probabilities, $\hat{x}_{\text{safe}}$ and $\hat{x}_{\text{alert}}$, derived by inverting the Bayesian belief functions at $\mu = \tau^*$:*

1. *High Suspicion Boundary ($\hat{x}_{\text{safe}}$): The honesty probability below which even a safe signal s_0 fails to reduce suspicion below τ^*. Solving $\mu(s_0) = \tau^*$ yields:*

$$\hat{x}_{\text{safe}} = \frac{(1-\eta_b)(1-\tau^*)}{(1-\eta_b)(1-\tau^*) + (1-\epsilon_b)\tau^*}.$$

2. *High Trust Boundary ($\hat{x}_{\text{alert}}$): The honesty probability above which even an alert signal s_1 is insufficient to raise suspicion above τ^*. Solving $\mu(s_1) = \tau^*$ yields:*

$$\hat{x}_{\text{alert}} = \frac{\eta_b(1-\tau^*)}{\eta_b(1-\tau^*) + \epsilon_b\tau^*}.$$

Given $\eta_b > \epsilon_b$, we have $\hat{x}_{\text{safe}} < \hat{x}_{\text{alert}}$. These boundaries define three extended audit regions, as shown in Table 2:

- *Region I (Indiscriminate escalation):* When $x < \hat{x}_{\text{safe}}$, the prior trust is so low that the auditor escalates regardless of the screening outcome, i.e., $\Gamma^* = (1, 1)$.
- *Region II (Conditional escalation):* When $\hat{x}_{\text{safe}} < x < \hat{x}_{\text{alert}}$, the auditor relies on the screening signal and escalates only after an alert, i.e., $\Gamma^* = (0, 1)$.
- *Region III (High trust):* When $x > \hat{x}_{\text{alert}}$, the cost of verification outweighs the perceived risk, leading to no escalation, i.e., $\Gamma^* = (0, 0)$.

5.3 Equilibrium Analysis in Region I

Region I is a high-suspicion region, where the auditor's posterior belief always exceeds the investigation threshold. Thus, the auditor executes the extended audit regardless of the screening outcome. The resulting mixed-strategy equilibrium is characterized as follows [Guo *et al.*, 2026b; Barron, 2024].

Theorem 1 (Audit-Intensive Equilibrium)**.** *Under the Region I escalation policy $\Gamma_{\text{I}} = (1, 1)$, the candidate strategy profile is $\Omega_{\text{I}} = (x_{\text{I}}^*, \alpha_{\text{I}}^*, \Gamma_{\text{I}})$, where*

$$\alpha_{\text{I}}^* = \frac{c_0 D_{\text{rem}} + \delta\left(A_1 z^{A_2 D_u} - A_3\right)}{V_{\text{pen}}(\eta_e - \epsilon_e)},$$

$$x_{\text{I}}^* = \frac{\eta_e(V_{\text{pen}} + \lambda_{\text{fn}}) - (c_b + c_e)}{\eta_e(V_{\text{pen}} + \lambda_{\text{fn}}) + \epsilon_e(\lambda_{\text{fp}} - V_{\text{pen}})}.$$

Here, $D_u = \sum_{i\in\mathcal{I}} d_i^u$ and $D_{\text{rem}} = \sum_{i\in\mathcal{I}}(d_i - d_i^u)$. If $0 \le x_{\text{I}}^* < \hat{x}_{\text{safe}}$ and $0 \le \alpha_{\text{I}}^* \le 1$, then Ω_{I} constitutes a feasible mixed-strategy equilibrium in Region I.

Remark: The proof of Theorem 1 is given in Appendix 2. In Region I, any execution of the basic audit always leads to the extended audit, regardless of the screening outcome. Consequently, the auditor's strategic decision reduces to how frequently to conduct a basic audit. The equilibrium reveals two key observations: i) lower auditing costs (c_b and c_e) increase the server's equilibrium honesty probability (higher x^*). ii) A larger conviction penalty increases the expected loss from being caught, which allows the auditor to inspect less frequently (lower α^*) while preserving deterrence.

5.4 Equilibrium Analysis in Region II

In Region II, an alert signal raises the posterior belief above the threshold and triggers escalation ($\gamma_1^* = 1$), while a safe signal suppresses escalation ($\gamma_0^* = 0$). Next, we characterize the equilibrium within this region in Theorem 2.

Theorem 2 (Audit-Light Equilibrium)**.** *Under the Region II escalation policy $\Gamma_{\text{II}} = (0, 1)$, the candidate strategy profile is $\Omega_{\text{II}} = (x_{\text{II}}^*, \alpha_{\text{II}}^*, \Gamma_{\text{II}})$, where*

$$\alpha_{\text{II}}^* = \frac{c_0 D_{\text{rem}} + \delta\left(A_1 z^{A_2 D_u} - A_3\right)}{V_{\text{pen}}(\eta_b\eta_e - \epsilon_b\epsilon_e)},$$

$$x_{\text{II}}^* = \frac{\eta_b[\eta_e(V_{\text{pen}} + \lambda_{\text{fn}}) - c_e] - c_b}{\eta_b[\eta_e(V_{\text{pen}} + \lambda_{\text{fn}}) - c_e] + \epsilon_b[c_e + \epsilon_e(\lambda_{\text{fp}} - V_{\text{pen}})]}.$$

If $\hat{x}_{\text{safe}} < x_{\text{II}}^* < \hat{x}_{\text{alert}}$ and $0 \le \alpha_{\text{II}}^* \le 1$, then Ω_{II} constitutes a feasible mixed-strategy equilibrium in Region II.

Remark: Theorem 2 is proved in Appendix 3. Region II characterizes the most effective operating regime of UAG. Here, the basic audit acts as a risk filter: alert signals trigger escalation, whereas safe signals terminate the audit. Instead of applying costly extended audits indiscriminately, the auditor escalates only when there is sufficient suspicion of noncompliance. This signal-dependent escalation makes screening quality crucial: improving screening accuracy enables the auditor to reduce the inspection frequency α^* while inducing a higher server honesty probability x^*.

5.5 Equilibrium Analysis in Region III

In Region III, even an alert signal fails to raise suspicion above τ^*, yielding a policy of zero escalation. However, we prove that no stable equilibrium exists in this region.

Proposition 1 (Non-Existence of Equilibrium in Region III)**.** *In Region III, no equilibrium can be sustained as the auditor's escalation threat is not credible.*

Algorithm 1 Equilibrium Characterization Algorithm

Input: c_0, c_b, c_e, z, V_{pen}, λ_{fn}, λ_{fp}, η_b, ϵ_b, η_e, ϵ_e, A_1, A_2, A_3, D_u, D_{rem}, δ.
Output: Equilibrium strategy set $\mathcal{E}^*$.
Compute $G_u \leftarrow c_0 D_{\text{rem}} + \delta(A_1 z^{A_2 D_u} - A_3)$.
Compute the investigation threshold:
    $\tau^* \leftarrow \frac{c_e+\epsilon_e(\lambda_{\text{fp}}-V_{\text{pen}})}{\eta_e(V_{\text{pen}}+\lambda_{\text{fn}})+\epsilon_e(\lambda_{\text{fp}}-V_{\text{pen}})}$.
Compute the strategic boundaries:
    $\hat{x}_{\text{safe}} \leftarrow \frac{(1-\eta_b)(1-\tau^*)}{(1-\eta_b)(1-\tau^*)+(1-\epsilon_b)\tau^*}$;
    $\hat{x}_{\text{alert}} \leftarrow \frac{\eta_b(1-\tau^*)}{\eta_b(1-\tau^*)+\epsilon_b\tau^*}$.
Initialize $\mathcal{E}^* \leftarrow \emptyset$.
Region I candidate equilibrium:
    $x_{\text{I}} \leftarrow \frac{\eta_e(V_{\text{pen}}+\lambda_{\text{fn}})-(c_b+c_e)}{\eta_e(V_{\text{pen}}+\lambda_{\text{fn}})+\epsilon_e(\lambda_{\text{fp}}-V_{\text{pen}})}$;
    $\alpha_{\text{I}} \leftarrow \frac{G_u}{V_{\text{pen}}(\eta_e-\epsilon_e)}$.
if $0 \le x_{\text{I}} < \hat{x}_{\text{safe}}$ and $0 \le \alpha_{\text{I}} \le 1$ **then**
    Add $\Omega_{\text{I}} = (x_{\text{I}}, \alpha_{\text{I}}, (1,1))$ to $\mathcal{E}^*$.
end if
Region II candidate equilibrium:
    $x_{\text{II}} \leftarrow \frac{\eta_b[\eta_e(V_{\text{pen}}+\lambda_{\text{fn}})-c_e]-c_b}{\eta_b[\eta_e(V_{\text{pen}}+\lambda_{\text{fn}})-c_e]+\epsilon_b[c_e+\epsilon_e(\lambda_{\text{fp}}-V_{\text{pen}})]}$;
    $\alpha_{\text{II}} \leftarrow \frac{G_u}{V_{\text{pen}}(\eta_b\eta_e-\epsilon_b\epsilon_e)}$.
if $\hat{x}_{\text{safe}} < x_{\text{II}} < \hat{x}_{\text{alert}}$ and $0 \le \alpha_{\text{II}} \le 1$ **then**
    Add $\Omega_{\text{II}} = (x_{\text{II}}, \alpha_{\text{II}}, (0,1))$ to $\mathcal{E}^*$.
end if
if $\mathcal{E}^* = \emptyset$ **then**
    Return NoFeasibleEquilibrium.
else
    Return $\mathcal{E}^*$.
end if

Remark: Proposition 1 is proved in Appendix 4. In a high-trust region, auditing appears economically irrational, as the expected benefit no longer offsets the audit cost, rendering the audit threat non-credible. However, once this auditing threat is removed, a rational server has an incentive to deviate to cheating to avoid the unlearning cost. Consequently, sustained compliance cannot be supported without a credible audit threat, causing the mechanism to collapse.

5.6 Equilibrium Characterization Algorithm

Based on the above equilibrium analysis, we develop the algorithm 1 for deriving equilibrium strategies. The algorithm first computes the investigation threshold τ^* and the honesty boundaries $\hat{x}_{\text{safe}}$ and $\hat{x}_{\text{alert}}$. It then evaluates the candidate equilibria for Regions I and II separately, retaining only those that satisfy the corresponding feasibility conditions. The algorithm returns the resulting set of feasible equilibrium strategies $\mathcal{E}^*$; if no candidate is feasible, it returns NoFeasibleEquilibrium.

6 Performance Evaluation

6.1 Experiment Setup

Models, Datasets & Settings. Our framework is evaluated on the CIFAR-10 dataset [Krizhevsky and others, 2009] using a ResNet-18 architecture. We consider a representative unlearning scenario with 100 data providers, with the training data uniformly distributed among them. To examine the impact of different unlearning intensities, the unlearning request size is varied within $[1, 300]$. The model accuracy degradation due to unlearning in Eq. (3) is modeled using an exponential decay function [Cui and Cheung, 2025] fitted from empirical results on a standard unlearning benchmark (i.e., SISA [Bourtoule *et al.*, 2021]), with parameters $A_1 = 10$, $A_2 = 0.01$, and $A_3 = 5$. The default conviction penalty is set to 400, while the liabilities for false negatives and false positives are set to 600 and 800, respectively.

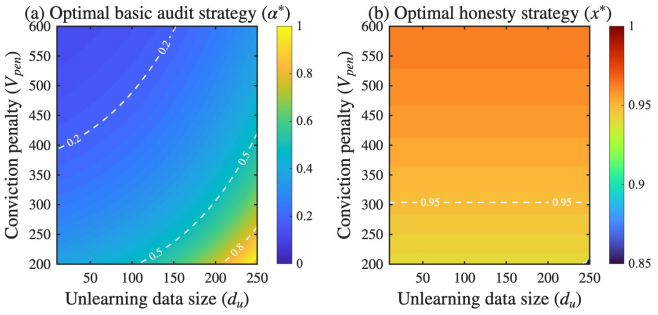

Figure 3: Equilibrium outcomes of our scheme under varying unlearning data size d_u and conviction penalty V_{pen}.

Comparative Methods. We compare UAG with two representative auditing approaches. (i) *Inference-based (MIA [Jia* et al.*, 2023]):* an efficient method that infers data membership by analyzing the confidence gap in the model's prediction outputs between training and non-training samples. (ii) *Reconstruction-based (TAPE [Wang* et al.*, 2025b]):* a high-precision audit approach by quantifying the posterior difference of model parameters to detect traces of erased data.

Evaluation Metrics. We evaluate system performance using five metrics: (i) *auditor utility (U_A)* and (ii) *server utility (U_S)*, which measure the expected payoffs of both players; (iii) *honesty ratio (x^*)*, defined as the equilibrium probability of the server faithfully executing the unlearning request; (iv) *audit reliability*, measured as the ratio of true positives to all positive audit outcomes. It indicates the trustworthiness of enforcement, representing the probability that a penalized server is indeed guilty; and (v) *cost-effectiveness ratio*, defined as the detection effectiveness normalized by the total operational and risk costs. Specifically, it captures how much effective deterrence is achieved per unit cost, accounting for both audit expenses and losses caused by misclassification.

6.2 Experimental Results

Analysis of Equilibrium Strategies. We first show that UAG induces a cost-effective equilibrium, enabling the auditor to maintain high server honesty with low audit frequency. Fig. 3 visualizes the joint impact of unlearning data size d_u and conviction penalty V_{pen} on the equilibrium strategies in Region II. As observed in Fig. 3 (a), the basic audit probability α^* adapts to the honesty probability. Specifically, as d_u increases, the server has stronger incentives to cheat, requiring a higher α^* to maintain deterrence. Conversely, increasing the penalty allows the auditor to reduce α^*, as stronger penalties substitute for frequent checks. Correspondingly, Fig. 3 (b) shows that UAG maintains robust compliance

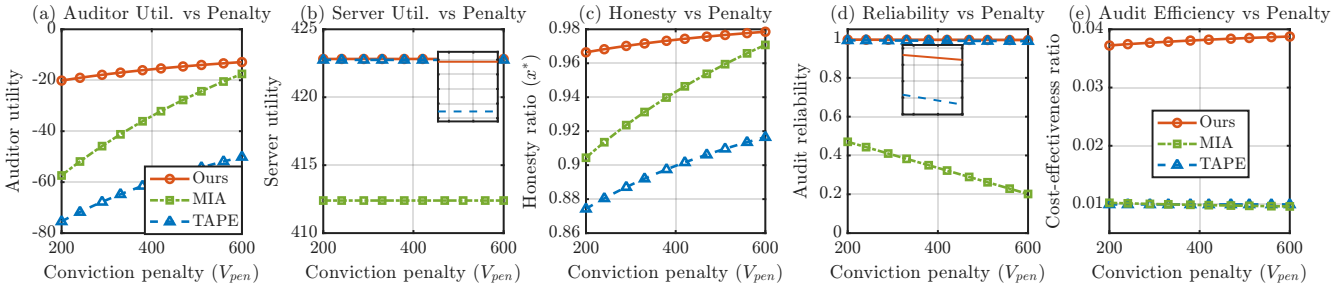

Figure 4: Performance comparison across three schemes under varying conviction penalty V_{pen}.

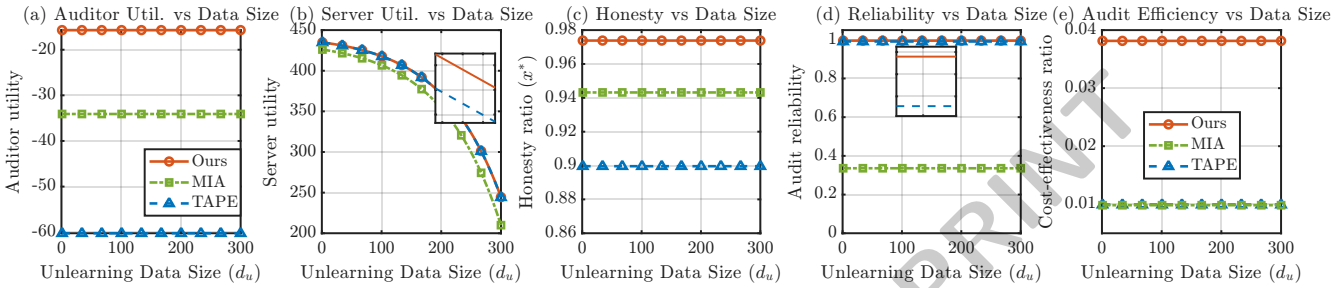

Figure 5: Performance comparison across three schemes under varying unlearning data size d_u.

(above 0.9) across most of the parameter space. Notably, under a moderate penalty (e.g., 300), an audit frequency of 0.5 suffices to enforce an honesty ratio close to 95%, demonstrating strong deterrence with limited auditing effort.

Impact of Conviction Penalty. UAG achieves strong deterrence under lower penalties while yielding higher utilities than baselines. Fig. 4 illustrates the system performance as the conviction penalty V_{pen} increases. *1) Higher utilities & effective deterrence:* Fig. 4 (a) and (b) show that UAG consistently outperforms MIA and TAPE in terms of payoffs. For the auditor, UAG avoids the high cost of always invoking the extended audit while mitigating the revenue loss from missed detections under MIA. For the server, utility remains high and stable because UAG reduces the risk of wrongful punishment via reliable verification. Consequently, as shown in Fig. 4 (c), UAG achieves a robust honesty ratio (above 0.96) with much lower penalties, whereas baselines require significantly harsher fines to induce comparable compliance. *2) Superior reliability & cost-effectiveness:* Fig. 4 (d) and (e) further highlight the operational advantages of UAG. Notably, in Fig. 4 (d), the reliability of MIA decreases as the penalty increases. This is because higher penalties increase the server's honesty, making actual cheating rare. Under such conditions, MIA's false alarms can dominate the few true detections, causing many penalized servers to be innocent. In contrast, UAG adopts a "screen-and-verify" design, where the basic audit flags suspicious cases and the extended audit confirms them. This two-stage verification effectively reduces wrongful punishment and maintains near-perfect audit reliability. Furthermore, Fig. 4 (e) shows that UAG achieves the highest cost-effectiveness ratio by reserving the expensive extended audit for only a small number of flagged cases, thereby providing TAPE-level reliability at a much lower cost.

Impact on Unlearning Data Size. UAG maintains stable audit performance and high server honesty even as the unlearning workload increases. Fig. 5 reports the system performance as unlearning request size d_u increases from 10 to 300. *1) Stable auditing performance:* As shown in Fig. 5 (a), (d), and (e), the auditor's utility, audit reliability, and efficiency remain invariant as d_u grows. It indicates that the auditing overhead in UAG is largely insensitive to the erased data size, consistent with TAPE, whose audit cost is dominated by the size of the user's local data rather than the volume of unlearned data. *2) Robust compliance incentives:* Fig. 5(b) shows that the server's utility decreases with d_u due to the increasing unlearning burden, yet UAG consistently achieves higher utility than the baselines. More importantly, Fig. 5 (c) shows the server's honesty ratio remains high and stable under UAG, demonstrating that UAG continues to incentivize honest unlearning even under heavy workloads.

7 Conclusion

In this paper, we address the key problem of enforcing truthful unlearning in MLaaS systems, where profit-driven servers may strategically misreport compliance. We propose **UAG**, a game-theoretic auditing framework that replaces exhaustive verification with strategic deterrence. To reduce auditing overhead, UAG adopts a hierarchical auditing mechanism that integrates low-cost screening with selectively triggered high-precision verification, and models the interaction between the auditor and the server as a dynamic Bayesian game. Through rigorous equilibrium analysis, we derive optimal auditing and compliance strategies that guide server behavior and deter deceptive unlearning. Extensive experiments demonstrate that UAG significantly reduces auditing overhead while maintaining a high level of server honesty.

Acknowledgments

This work was supported in part by NSFC (Nos. U24A20237, U22A2029, 62302387, U23A20276) and the Youth Innovation Team of Shaanxi Universities.

References

[Barron, 2024] Emmanual N Barron. *Game theory: an introduction*. John Wiley & Sons, 2024.

[Bourtoule *et al.*, 2021] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2021.

[Chen *et al.*, 2021] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 896–911, 2021.

[Cui and Cheung, 2025] Yue Cui and Man Hon Cheung. The Price of Forgetting: Incentive Mechanism Design for Machine Unlearning. *IEEE Transactions on Mobile Computing*, 24(11):11852–11864, 2025.

[Golatkar *et al.*, 2021] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-Privacy Forgetting in Deep Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 792–801, 2021.

[Guo *et al.*, 2024] Yu Guo, Yu Zhao, Saihui Hou, Cong Wang, and Xiaohua Jia. Verifying in the Dark: Verifiable Machine Unlearning by Using Invisible Backdoor Triggers. *IEEE Transactions on Information Forensics and Security*, 19:708–721, 2024.

[Guo *et al.*, 2025] Shaolong Guo, Yuntao Wang, Ning Zhang, Zhou Su, Tom H. Luan, Zhiyi Tian, and Xuemin Shen. A survey on semantic communication networks: Architecture, security, and privacy. *IEEE Communications Surveys Tutorials*, 27(5):2860–2894, 2025.

[Guo *et al.*, 2026a] Shaolong Guo, Yuntao Wang, Zhou Su, and Tom H. Luan. Incentivizing Truthful Machine Unlearning via Hierarchical Auditing: Online appendix. [Online]. Available: https://drive.google.com/file/d/1Qln6RlVKcI8nWYIjvys8tPBYcbC8ut9R/view?usp=sharing, 2026.

[Guo *et al.*, 2026b] Shaolong Guo, Yuntao Wang, Zhou Su, Yanghe Pan, Tom H. Luan, and Xizhao Luo. FLET: Game-theoretic free-riding mitigation via test tasks in federated learning. *IEEE Transactions on Networking*, 34:3995–4010, 2026.

[Han *et al.*, 2025] Mengde Han, Tianqing Zhu, Lefeng Zhang, Huan Huo, and Wanlei Zhou. Vertical federated unlearning via backdoor certification. *IEEE Transactions on Services Computing*, 18(2):1110–1123, 2025.

[He *et al.*, 2025] Zhengbao He, Tao Li, Xinwen Cheng, Zhehao Huang, and Xiaolin Huang. Towards natural machine unlearning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11548–11560, 2025.

[Hu *et al.*, 2022] Hongsheng Hu, Zoran Salčić, Gillian Dobbie, Jinjun Chen, Lichao Sun, and Xuyun Zhang. Membership inference via backdooring. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3832–3838, 2022.

[Jia *et al.*, 2023] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, volume 36, pages 51584–51605, 2023.

[Jiao *et al.*, 2021] Yutao Jiao, Ping Wang, Dusit Niyato, Bin Lin, and Dong In Kim. Toward an automated auction framework for wireless federated learning services market. *IEEE Transactions on Mobile Computing*, 20(10):3034–3048, 2021.

[Krizhevsky and others, 2009] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *Technical Report*, pages 1–60, 2009.

[Kurmanji *et al.*, 2023] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, volume 36, pages 1957–1987, 2023.

[Lee *et al.*, 2025] Hong kyu Lee, Qiuchen Zhang, Carl Yang, Jian Lou, and Li Xiong. Contrastive unlearning: A contrastive approach to machine unlearning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7464–7472, 2025.

[Li *et al.*, 2025] Zitong Li, Qingqing Ye, and Haibo Hu. Funu: Boosting machine unlearning efficiency by filtering unnecessary unlearning. In *Proceedings of the ACM on Web Conference (WWW)*, pages 3366–3376, 2025.

[Lin *et al.*, 2024] Yijing Lin, Zhipeng Gao, Hongyang Du, Dusit Niyato, Jiawen Kang, and Xiaoyuan Liu. Incentive and Dynamic Client Selection for Federated Unlearning. In *Proceedings of the ACM on Web Conference (WWW)*, pages 2936–2944, 2024.

[Liu *et al.*, 2025] Hengzhu Liu, Tianqing Zhu, Lefeng Zhang, and Ping Xiong. Game-Theoretic Machine Unlearning: Mitigating Extra Privacy Leakage. *IEEE Transactions on Information Forensics and Security*, 20:11591–11606, 2025.

[Lyu *et al.*, 2025] Hongyi Lyu, Xuyun Zhang, Hongsheng Hu, Shuo Wang, Chaoxiang He, and Lianyong Qi. Fine-grained and efficient self-unlearning with layered iteration. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7643–7651, 2025.

[Pan *et al.*, 2026] Yanghe Pan, Yuntao Wang, Zhou Su, Yuan Gao, Shaolong Guo, Qinnan Hu, Ruidong Li, and Wei-Wei Li. The right to be forgotten versus the need to be

remembered: Efficient personalized federated unlearning with optimal incentives. *IEEE Transactions on Network Science and Engineering*, 13:1616–1631, 2026.

[Sommer *et al.*, 2022] David M. Sommer, Liwei Song, Sameer Wagh, and Prateek Mittal. Athena: Probabilistic verification of machine unlearning. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, (3):268–290, 2022.

[Wang *et al.*, 2021] Yuntao Wang, Zhou Su, Ning Zhang, and Abderrahim Benslimane. Learning in the air: Secure federated learning for uav-assisted crowdsensing. *IEEE Transactions on Network Science and Engineering*, 8(2):1055–1069, 2021.

[Wang *et al.*, 2022] Yuntao Wang, Weiwei Chen, Tom H. Luan, Zhou Su, Qichao Xu, Ruidong Li, and Nan Chen. Task offloading for post-disaster rescue in unmanned aerial vehicles networks. *IEEE/ACM Transactions on Networking*, 30(4):1525–1539, 2022.

[Wang *et al.*, 2024] Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Machine unlearning: A comprehensive survey. *arXiv preprint arXiv:2405.07406*, pages 1–33, 2024.

[Wang *et al.*, 2025a] Qiyuan Wang, Ruiling Xu, Shibo He, Randall Berry, and Meng Zhang. Unlearning Incentivizes Learning under Privacy Risk. In *Proceedings of the ACM on Web Conference (WWW)*, pages 1456–1467, 2025.

[Wang *et al.*, 2025b] Weiqi Wang, Zhiyi Tian, An Liu, and Shui Yu. TAPE: Tailored posterior difference for auditing of machine unlearning. In *Proceedings of the ACM on Web Conference (WWW)*, page 3061–3072, 2025.

[Warnecke *et al.*, 2023] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *Proceedings of the 30th Network and Distributed System Security (NDSS)*, pages 1–18, 2023.

[Weng *et al.*, 2024] Jiasi Weng, Shenglong Yao, Yuefeng Du, Junjie Huang, Jian Weng, and Cong Wang. Proof of Unlearning: Definitions and Instantiation. *IEEE Transactions on Information Forensics and Security*, 19:3309–3323, 2024.

[Wu and Harandi, 2025] Jing Wu and Mehrtash Harandi. MUNBa: Machine unlearning via nash bargaining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4754–4765, 2025.

[Xu *et al.*, 2026] Heng Xu, Tianqing Zhu, Lefeng Zhang, and Wanlei Zhou. Really unlearned? verifying machine unlearning via influential sample pairs. *IEEE Transactions on Dependable and Secure Computing*, 23(01):1671–1686, 2026.

[Yang *et al.*, 2025] Zhe-Rui Yang, Jindong Han, Chang-Dong Wang, and Hao Liu. Erase then rectify: A training-free parameter editing approach for cost-effective graph unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 13044–13051, 2025.

[Zhang *et al.*, 2024] Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. Verification of machine unlearning is fragile. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 58717–58738, 2024.