

Machine Unlearning: Techniques, Challenges, Evaluation, Applications, and Future Directions

Akshat Kumawat^{1*}, Balraj², Dr. Vishal Shrivastava³, Dr. Akhil Pandey⁴

^{1,2,3,4} Department of Artificial Intelligence & Data Science, Arya College of Engineering & I.T., Jaipur, India

Abstract—The widespread deployment of deep neural networks and foundation models has introduced acute legal, operational, and ethical challenges regarding data privacy and algorithmic accountability. Regulatory mandates, notably the European Union General Data Protection Regulation (GDPR Article 17 "Right to be Forgotten") and the California Consumer Privacy Act (CCPA), legally obligate data controllers to purge personal records upon individual request. However, physical data deletion from storage disks fails to sanitize trained models: gradient optimization permanently embeds training data signatures into overparameterized weights. While complete model retraining from scratch on retained data provides exact statistical revocation, its prohibitive computational, financial, and environmental costs render it intractable for modern architectures and streaming deletion requests. To overcome this bottleneck, Machine Unlearning has emerged as a fundamental discipline in trustworthy artificial intelligence, aiming to excise the influence of designated training subsets without compromising general model utility on retained data. This paper presents an academically rigorous, comprehensive survey of machine unlearning. We mathematically formalize the unlearning problem under Empirical Risk Minimization (ERM), classify paradigms into exact, approximate, and certified regimes, and establish an exhaustive taxonomy spanning data-partitioning (SISA), gradient inversion, influence functions, knowledge distillation, parameter-saliency pruning, certified differential perturbations, federated unlearning, and autoregressive Large Language Models (LLMs). We examine multi-dimensional evaluation suites—including membership inference attacks (MIA), loss distribution divergence, and relearning susceptibility protocols—while systematically contrasting unlearning with differential privacy and model editing. Finally, we analyze critical vulnerabilities and outline thirteen structured future research directions for verifiable unlearning systems.

Keywords: Machine Unlearning, Machine Learning, Data Privacy, Right to be Forgotten, Certified Unlearning, Differential Privacy, Federated Learning, Large Language Models.

1. INTRODUCTION

1.1 Background

Over the past decade, machine learning (ML) has transitioned from localized statistical modeling into the operational infrastructure of global digital systems. Deep neural networks (DNNs), vision transformers, and foundation models are trained on massive datasets scraped from the public web, enterprise archives, and clinical databases [1], [2]. While predictive capabilities scale directly with parameter capacity and dataset volume according to empirical scaling laws [41], [42], overparameterized networks generalize not only by learning abstract invariants, but also by memorizing idiosyncratic training instances, verbatim text sequences, and personally identifiable information (PII) [17], [18]. When deployed in high-stakes domains—such as clinical diagnostics, automated lending, and biometric surveillance—memorization creates acute privacy vulnerabilities [18], [19].

1.2 Data Privacy Problem in Trained Machine Learning Models

Organizations regularly face statutory, ethical, and commercial obligations to revoke data from operational models. Unlike relational databases (RDBMS) where an UPDATE or DELETE command physically erases records from disk, deep learning models store knowledge in a distributed, non-localized manner across continuous parameter arrays [1], [5]. Sensitive contexts necessitating post-training removal include:

- **Personally Identifiable Information (PII):** Residential addresses, phone numbers, and credentials ingested during web crawls [17], [19].
- **Clinical Records:** Patient diagnostic histories, radiological scans, and genomic sequences protected under HIPAA and GDPR [35].
- **Financial Histories:** Transaction ledgers, credit histories, and proprietary trading strategies [36].

- **User-Generated Content:** Media and messages uploaded by users who subsequently revoke consent or terminate accounts [2], [15].
- **Copyrighted and Proprietary Assets:** Licensed literary works, code repositories, and artwork ingested without sustained permissions [24], [33].
- **Adversarial Contaminants:** Data poisoning instances, backdoors, and defamatory falsehoods [27], [30].

When adversaries query models containing such data via black-box APIs or white-box inspection, extraction and inference attacks can reconstruct these sensitive instances with high fidelity [17], [18].

1.3 The Right to be Forgotten and Regulatory Mandates

In response to pervasive data harvesting, global regulatory frameworks have codified strict privacy rights. The European Union's General Data Protection Regulation (GDPR) establishes the "Right to Erasure" under Article 17, colloquially known as the Right to be Forgotten (RtbF). Article 17 empowers individuals to demand the permanent deletion of personal data when consent is withdrawn or processing is unlawful [35], [37]. Parallel statutes include the California Consumer Privacy Act (CCPA / CPRA Section 1798.105) and Canada's PIPEDA. Crucially, regulatory authorities interpret "processing" to encompass derivative predictive models. A model retaining the ability to infer or output private data is legally considered an unauthorized data store [2], [35], exposing operators to administrative fines up to €20 million or 4% of global annual turnover under GDPR Article 83.

1.4 Motivation for Machine Unlearning

Physical data deletion does not sanitize trained neural networks. Purging records from database clusters leaves parameter weights completely unchanged. Querying adversaries can execute Membership Inference Attacks (MIAs) [18], exploiting confidence distributions to establish that the purged record participated in model training. In foundation models, prompt elicitation can extract verbatim sequences [17].

The only mathematically exact baseline within standard learning theory is Complete Retraining from Scratch: discarding the target forget set D_f from primary dataset D , randomly reinitializing parameters θ_0 , and re-optimizing exclusively over the retain set $D_r = D \setminus D_f$. While retraining provides a gold-standard guarantee—since retrained model M_r has zero statistical exposure to D_f —it is computationally, economically, and environmentally prohibitive. Training modern deep networks demands thousands of GPU hours, millions of dollars, and substantial carbon emissions [2], [43]. Under continuous, streaming deletion requests, retraining from scratch creates an impossible computational backlog [2], [38]. Machine unlearning addresses this challenge by providing post-hoc algorithmic transformations that replicate the statistical properties of retraining at a fraction of the computational cost.

1.5 Central Research Problem Formulation

Formally, the central research problem of Machine Unlearning is formulated as:

Definition	1:	The	Central	Machine	Unlearning	Problem
<i>Given a complete dataset D, a trained model M parameterized by $\theta \in \mathbb{R}^d$, a designated forget subset $D_f \subset D$, and a retain subset $D_r = D \setminus D_f$, design an unlearning operator $U: (M, D, D_f, D_r) \rightarrow M_u$ such that the unlearned model M_u satisfies two simultaneous objectives:</i>						
<i>1. Efficacy and Privacy: The functional behavior and parameter distribution of M_u are statistically indistinguishable from a reference model M_r trained exclusively on D_r without exposure to D_f: $\text{Dist}(P(M_u), P(M_r)) \leq \epsilon$.</i>						
<i>2. Utility Preservation and Efficiency: M_u preserves predictive accuracy on D_r and general test distributions, while the computational complexity and wall-clock execution time of U are substantially lower than full retraining: $\text{Cost}(U) \ll \text{Cost}(\text{Retrain}(D_r))$.</i>						

1.6 Contributions of this Paper

This paper provides an authoritative, mathematically rigorous synthesis of machine unlearning across the following core dimensions:

- **1. Methodological Taxonomy:** We construct an exhaustive taxonomy categorizing unlearning techniques into exact sharding (SISA), first-order gradient reversal, second-order influence approximations, relational knowledge distillation, parameter-saliency masking, certified privacy perturbations, federated unlearning, and foundation model unlearning.

- **2. Theoretical Formulations:** We formalize Empirical Risk Minimization (ERM), exact versus approximate bounds, parameter-distance limits, and (ϵ, δ) -certified unlearning guarantees under formal privacy frameworks.
- **3. Multi-Dimensional Evaluation Framework:** We systematically review evaluation methodologies, showing why forget-set accuracy is fundamentally insufficient and establishing the necessity of membership inference auditing, loss-distribution divergence, and relearning susceptibility protocols.
- **4. Comparative Paradigm Synthesis:** Through structured comparative tables, we contrast unlearning against complete retraining, physical data deletion, differential privacy, and knowledge editing across performance, privacy, and compute axes.
- **5. Application Domains:** We analyze real-world operational requirements across healthcare, finance, recommendation systems, edge federated systems, computer vision, and foundation models.
- **6. Critical Analysis and Future Roadmap:** We identify core vulnerabilities (such as decision boundary masking, catastrophic utility collapse, and sequential degradation) and formulate thirteen structured research directions to guide future scholarship.

2. BACKGROUND AND FUNDAMENTALS

2.1 Machine Learning Training Paradigms

Supervised machine learning maps an input space $X \subseteq \mathbb{R}^p$ to a label space Y . A training dataset $D = \{(x_i, y_i)\}_{i=1}^N$ comprises N i.i.d. observations drawn from data-generating distribution $P_{X \times Y}$. A parametric model family $f_\theta: X \rightarrow Y$ is governed by parameters $\theta \in \Theta \subseteq \mathbb{R}^d$. Discrepancies between predictions $f_\theta(x)$ and ground-truth labels y are quantified by loss function $\ell(f_\theta(x), y)$. For K -class categorical classification, the standard cross-entropy loss is given by:

$$\begin{aligned} \ell_{\text{CE}}(f_\theta(x), y) &= - \sum_{k=1}^K I(y = k) \cdot \ln \left(\frac{\exp(z_k(x; \theta))}{\sum_{j=1}^K \exp(z_j(x; \theta))} \right) \end{aligned} \quad (1)$$

where $z(x; \theta) \in \mathbb{R}^K$ represents the unnormalized logit outputs of the neural network. Under Empirical Risk Minimization (ERM), optimal parameters θ^* minimize the empirical risk over dataset D alongside a regularization penalty $\Omega(\theta)$:

$$\begin{aligned} L(\theta; D) &= (1 / |D|) \sum_{(x_i, y_i) \in D} \ell(f_\theta(x_i), y_i) + (\lambda / 2) \|\theta\|_2^2, \quad \theta^* \\ &= \operatorname{argmin}_{\theta \in \Theta} L(\theta; D) \end{aligned} \quad (2)$$

Iterative optimization proceeds via mini-batch Stochastic Gradient Descent (SGD):

$$\theta_{t+1} = \theta_t - \eta_t \left[(1 / B) \sum_{(x_i, y_i) \in B_t} \nabla_\theta \ell(f_\theta(x_i), y_i) + \lambda \theta_t \right] \quad (3)$$

where $B_t \subset D$ is a mini-batch of size B , and $\eta_t > 0$ is the learning rate.

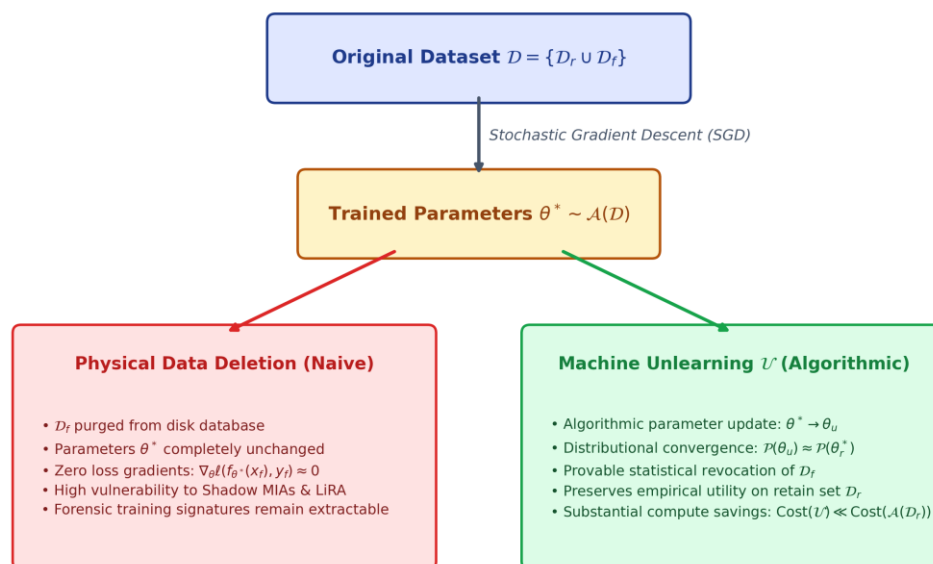


Fig. 1: Machine Learning Training vs. Machine Unlearning Paradigm. Physical data deletion (left) purges records from disk storage but leaves model weights θ^* untouched, whereas machine unlearning (right) algorithmically transforms parameters to produce $\theta_u \approx \theta_r^*$.

2.2 Training Data and Model Knowledge Encoding

Overparameterized neural networks function as distributed associative memories rather than compact functional compressions [17]. Information is encoded across deep layers: (1) Feature Extraction Manifolds, where early layers capture textures, syntactic tokens, and spatial primitives; (2) Latent Representations, where intermediate layers cluster instances, allocating specific activation directions to outliers; and (3) Decision Boundaries, where final projection layers position separating hyperplanes warped by individual training instances [2], [5]. Every weight parameter is the cumulative product of stochastic gradient projections induced by all observed training samples.

2.3 Physical Data Deletion versus Model Parameter Deletion

Equating database record deletion with model sanitization is a fundamental privacy fallacy. Deleting record r_k from disk returns NULL on queries, but leaves the trained model binary completely untouched. When queried with x_k , the model produces low-entropy predictions $f_{\theta}(x_k) \approx y_k$ and vanishing loss gradients $\nabla_{\theta} \ell(f_{\theta}(x_k), y_k) \approx 0$, acting as a persistent forensic record of the deleted data [18], [19]. True deletion requires modifying parameter weights to revoke this statistical dependency.

2.4 Privacy Vulnerabilities and Empirical Memorization

Deep networks readily memorize arbitrary patterns without generalizable structure [1]. Large Language Models memorize verbatim sequences from pre-training corpora—including PII and medical records—that can be extracted via prefix prompting [17]. Furthermore, neural networks are chronically vulnerable to Membership Inference Attacks (MIAs) [18], [19], where adversaries exploit prediction confidence and loss distributions to determine whether a target record was used in training, directly compromising confidentiality.

2.5 Conceptual Introduction to Machine Unlearning

Introduced by Cao and Yang in 2015 [1], machine unlearning investigates algorithmic protocols to excise the influence of designated data subsets without full retraining. While early methods addressed linear classifiers via statistical summation [1], modern unlearning focuses on deep neural networks. Here, non-convex loss surfaces imply path-dependent, stochastic optimization trajectories, meaning that simply subtracting historical gradients fails to restore the retrained model state [2], [38].

3. MACHINE UNLEARNING PROBLEM DEFINITION

3.1 Formal Mathematical Formulation and Set Partitioning

Let $Z = X \times Y$. The complete training dataset is $D = \{z_i\}_{i=1}^N \sim P_Z$. Upon a deletion request, D is partitioned into two mutually disjoint subsets:

$$D = D_r \cup D_f, \quad D_r \cap D_f = \emptyset \tag{4}$$

where $D_f = \{z_j\}_{j=1}^{N_f}$ is the forget set and $D_r = \{z_k\}_{k=1}^{N_r}$ is the retain set ($N_f \ll N_r$). Applying stochastic training algorithm $A: P(Z) \rightarrow \Theta$ to D yields original parameters $\theta^* \sim A(D)$. Applying A exclusively to the retain partition yields the retrained reference parameters: $\theta_{r^*} \sim A(D_r)$. The model $M_r = f_{\theta_{r^*}}$ represents the gold standard: its mutual information with D_f given D_r is strictly zero:

$$I(\theta_{r^*}; D_f \mid D_r) = 0 \tag{5}$$

An unlearning algorithm U produces unlearned parameters $\theta_u \sim U(\theta^*, D_f, D_r)$.

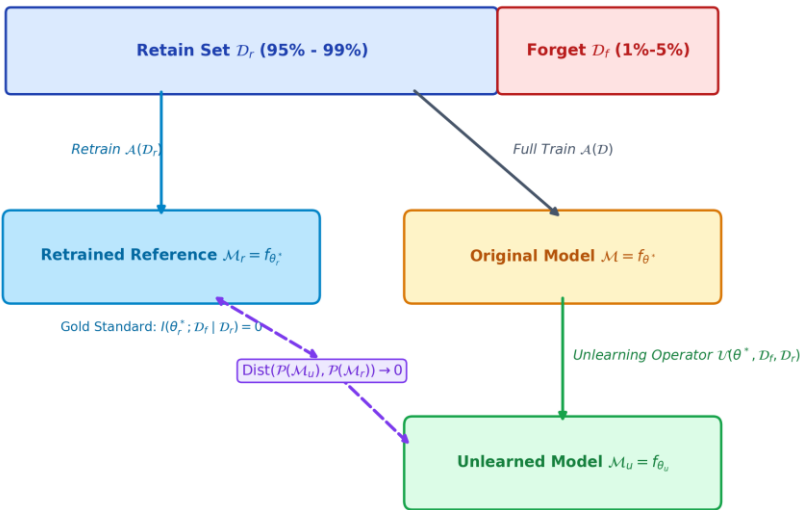


Fig. 2: Partitioning and Parameter Convergence in Machine Unlearning. The unlearned model μ_u must achieve distributional convergence with the retrained reference μ_r .

Table 1: Core Machine Unlearning Terminology and Mathematical Notations

Symbol / Term	Formal Definition	Operational Significance
D	Original complete dataset: $D = \{(x_i, y_i)\}_{i=1}^N$	Dataset used to train original deployed model M .
D_f	Target forget set: $D_f \subset D, D_f = N_f$	Sensitive samples requested for complete erasure.
D_r	Authorized retain set: $D_r = D \setminus D_f, D_r = N_r$	Authorized data on which model utility must be preserved.
$A(\cdot)$	Stochastic training pipeline: $P(Z) \rightarrow \Theta$	Base learning algorithm (SGD/Adam) mapping data to parameters.
$M = f_{\theta^*}$	Original trained model: $\theta^* \sim A(D)$	Initial model containing embedded influence of all data.
$M_r = f_{\theta_{r^*}}$	Retrained reference model: $\theta_{r^*} \sim A(D_r)$	Gold standard baseline: trained with zero exposure to D_f .
$\mu_u = f_{\theta_u}$	Unlearned sanitized model: $\theta_u \sim U(\theta^*, D_f, D_r)$	Output of unlearning algorithm U ; must approximate M_r .
(ϵ, δ)	Certified privacy bounds in hypothesis testing	Maximum permissible distribution discrepancy between μ_u and M_r .
H_r	Empirical Hessian on retain data: $\nabla_{\theta^2}^2 L(\theta^*; D_r)$	Second-order loss curvature governing parameter sensitivity.

3.2 The Unlearning Objective and Distributional Equivalence

The operational objective is that unlearned model μ behaves identically to retrained model μ_r : $\mu \approx \mu_r$. Because training algorithm A and unlearning operator U are stochastic processes, equivalence is formally defined at the distributional level: $P(\theta_u) \approx P(\theta_{r^*})$. Functionally, for any test query $x \in X$ and measurable predicate S :

$$| P(f_{\theta_u}(x) \in S) - P(f_{\theta_{r^*}}(x) \in S) | \leq \gamma, \quad \gamma \geq 0$$

(6)

3.3 Theoretical Hierarchy: Exact, Approximate, and Certified Unlearning

Machine unlearning techniques fall into three rigorous theoretical regimes:

- 1. **Exact Machine Unlearning:** Guarantees identical parameter distributions: $P(U(A(D), D_f, D_r)) \equiv P(A(D_r))$. Provides perfect privacy, but in deep networks requires architectural data partitioning (e.g., SISA) [2], [3].
- 2. **Approximate Machine Unlearning:** Relaxes strict equality to bounded statistical distance: $\text{Cost}(U) \ll \text{Cost}(A(D_r))$ and $D_{\text{eval}}(\mu, \mu_r) \leq \epsilon_{\text{tol}}$ [5], [12]. Delivers high speedups ($5\times\text{--}100\times$), but risks residual latent traces.
- 3. **Certified Machine Unlearning:** Injects calibrated noise into parameter updates to provide formal statistical indistinguishability [4], [7], [8].

Definition 2: Certified (ϵ, δ) -Unlearning (Guo et al., 2020; Sekhari et al., 2021)
An unlearning algorithm U provides (ϵ, δ) -certified unlearning if for all datasets D , forget sets $D_f \subset D$, and all measurable hypothesis subsets $S \subseteq \Theta$:
$$P(U(A(D), D_f) \in S) \leq \exp(\epsilon) \cdot P(A(D \setminus D_f) \in S) + \delta$$

and symmetrically, bounding an adversary's hypothesis testing advantage under arbitrary prior knowledge.

4. TAXONOMY OF MACHINE UNLEARNING TECHNIQUES

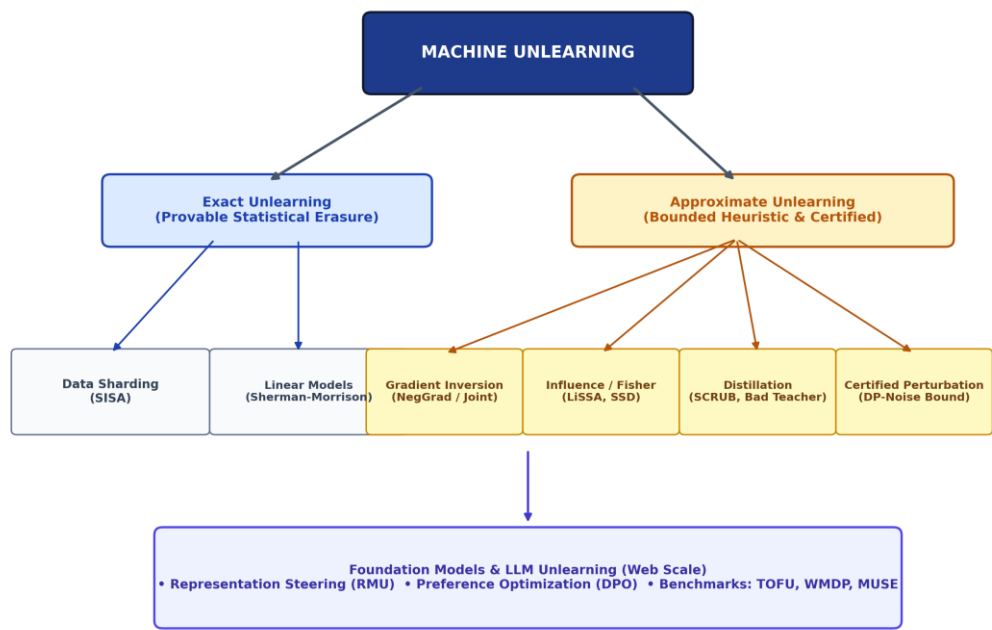


Fig. 3: Comprehensive taxonomy of machine unlearning techniques spanning exact, approximate, certified, and foundation model paradigms.

4.1 Exact Machine Unlearning

Exact unlearning eliminates all statistical dependence on D_f . In strictly convex models (ridge regression, linear models), cached covariance matrices $X^T X$ and target projections $X^T y$ allow vector subtraction: $(X_{-r}^T X_{-r}) = (X^T X) - x_f x_f^T$, and $(X_{-r}^T y_{-r}) = (X^T y) - x_f y_f$. The inverse covariance matrix is updated via the Sherman-Morrison rank-1 formula:

$$(X_{-r}^T X_{-r})^{-1} = (X^T X)^{-1} + [(X^T X)^{-1} x_f x_f^T (X^T X)^{-1}] / [1 - x_f^T (X^T X)^{-1} x_f] \quad (7)$$

This computes exact updated weights $\theta_u = (X_{-r}^T X_{-r})^{-1} (X_{-r}^T y_{-r})$ in $O(d^2)$ operations rather than $O(d^3)$ full matrix inversion [1], [3]. In non-convex deep networks, exact unlearning requires architectural data partitioning [2].

4.2 Approximate Machine Unlearning Principles

Approximate unlearning executes bounded optimization directly on pre-trained weights $\theta^* \rightarrow \theta_u \approx \theta_{r^*}$, providing $5\times-100\times$ speedups over retraining. However, it introduces critical trade-offs: (1) decision boundary masking, where output labels shift without dissolving internal feature representations, and (2) catastrophic utility collapse, where unlearning damages retain accuracy [12], [14].

4.3 Data Partitioning and Sharding (SISA)

Partitioning restricts sample influence architecturally. Dataset D is divided into S disjoint shards $D = \bigcup_{s=1}^S D^{(s)}$, training independent models $M^{(s)}$ aggregated via ensemble averaging:

$$f_{\text{ens}}(x) = (1/S) \sum_{s=1}^S f(x; \theta^{(s)}) \quad (8)$$

When record $z_f \in D^{(s^*)}$ is unlearned, only sub-model $M^{(s^*)}$ is retrained on $D^{(s^*)} \setminus \{z_f\}$, saving compute by factor S [2].

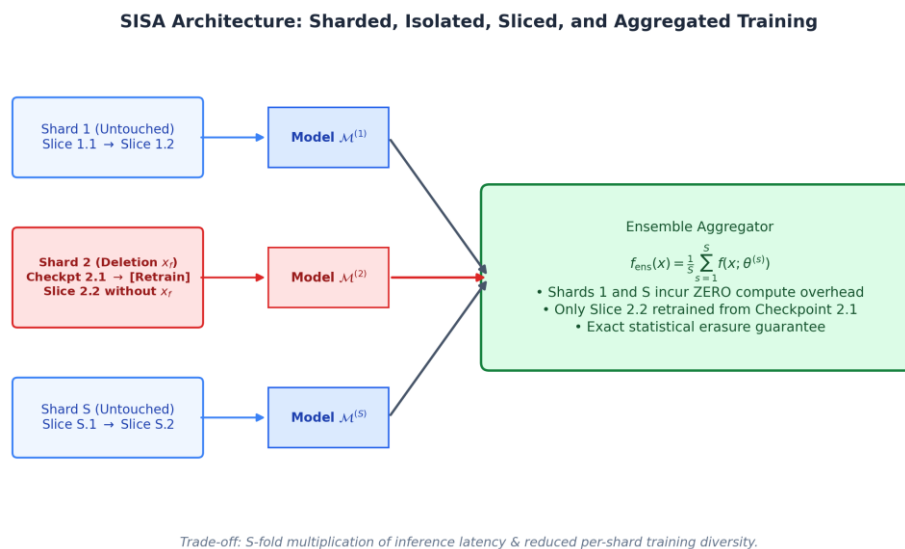


Fig. 4: SISA framework: Sharded, Isolated, Sliced, and Aggregated training. Deletion of record x_f in Shard 2 requires retraining only that shard from its checkpoint, leaving all other shards untouched.

4.4 SISA-Based Approaches

The SISA framework [2] combines Sharding (disjoint data subsets) with Slicing (chronological training increments with saved checkpoints $\theta^{(s, r)}$). To unlearn a record in slice r^* , SISA loads checkpoint $\theta^{(s^*, r^*-1)}$ and retrains exclusively over subsequent slices without z_f . While providing exact unlearning guarantees and model-agnosticism, SISA multiplies inference costs S -fold and reduces sample diversity per shard [2], [38].

4.5 Gradient-Based Unlearning

Gradient-based unlearning inverts optimization on D_f [5], [8], [31]. Negative Gradient Ascent (NegGrad) reverses loss gradients: $\theta_{t+1} = \theta_t + \eta \nabla_{\theta} L(\theta_t; D_f)$. Because unconstrained ascent pushes parameters into degenerate loss regions, modern formulations use regularized joint objectives:

$$\min_{\theta} [L(\theta; D_r) - \alpha L(\theta; D_f) + \beta \|\theta - \theta^*\|_2^2] \quad (9)$$

balancing retain utility, forget loss maximization, and weight proximity to preserve stability [8], [12].

4.6 Influence-Function-Based Approaches

Influence functions estimate parameter shifts via second-order Taylor expansions around minimizer θ^* [29]: $I_{\text{param}}(z_f) = -H_{\theta}^{-1} \nabla_{\theta} \ell(z_f, \theta)$. Removing z_f yields a single Newton-step correction:

$$\theta_u \approx \hat{\theta} + (1/N) H_{\theta}^{-1} \nabla_{\theta} \ell(z_f, \hat{\theta}) \quad (10)$$

While executing in a single step, explicitly calculating and inverting the Hessian matrix $H \in \mathbb{R}^{d \times d}$ is computationally intractable for modern networks where $d > 10^7$, requiring stochastic approximations such as LiSSA (Linear time Stochastic Second-Order Algorithm) [29], [30]. Furthermore, non-convexity and negative eigenvalues in deep networks can cause severe numerical instability [38].

4.7 Knowledge Distillation Frameworks

Distillation purges information via asymmetric teacher-student interactions [6], [10], [11], [12]. Incompetent Teacher Unlearning [11] forces student M_u to mimic competent teacher M_{comp} on D_r while mimicking uniform random distributions from incompetent teacher M_{bad} on D_f . SCRUB [12] formulates unlearning as a min-max game, maximizing divergence on D_f while bounding representation distance on D_r to prevent utility collapse.

4.8 Parameter-Based Unlearning and Saliency Pruning

Parameter-based methods localize unlearning to sparse weight subsets [13], [14], [31], [32]. Selective Synaptic Dampening (SSD) [13] computes diagonal Fisher Information F_D and F_{D_f} , dampening weights that encode D_f without contributing to retain utility:

$$\theta_k^{\text{unlearn}} = \theta_k^* \cdot \min(1, [F_k(D_r)] / [\alpha F_k(D_f) + \gamma]) \quad (11)$$

SalUn [14] computes gradient saliency $S_k = |\partial L(\theta; D_f) / \partial \theta_k|$, projecting updates onto the top 5% salient weights while freezing the remaining 95%, preventing collateral retain damage.

4.9 Certified Machine Unlearning and Differential Perturbation

Certified unlearning adds calibrated Gaussian noise $b \sim N(0, \sigma^2 I)$ to analytical parameter updates [4], [7], [8]: $\theta_u = \theta^* + H_{\theta}^{-1} \nabla_{\theta} L(\theta^*; D_f) + b$, scaling the noise standard deviation σ to operator sensitivity:

$$\sigma \geq [\Delta_2 \sqrt{2 \ln(1.25 / \delta)}] / \epsilon \quad (12)$$

This guarantees that adversaries cannot infer training set membership beyond theoretical (ϵ, δ) hypothesis testing bounds [4], [8].

4.10 Federated Machine Unlearning

In Federated Learning (FL), unlearning targets: (1) Sample-Level Unlearning, removing local client data, and (2) Client-Level Unlearning, excising all historical update vectors from departing edge clients [15], [16]. FedEraser [15] reconstructs global trajectories by projecting historical updates without raw data re-transmission. FedRecovery [16] combines gradient caching with DP noise for rapid client removal.

4.11 Machine Unlearning for Large Language Models

LLMs present unique unlearning hurdles due to continuous autoregressive sequence generation $P(w_{1:T}) = \prod_{t=1}^T P(w_t | w_{<t})$ [21], [24], [33], [40]. Unlearning targets: (1) factual triples [22], [23]; (2) personal profiles (TOFU benchmark) [21]; (3) hazardous CBRN knowledge (WMDP benchmark / RMU) [24]; and (4) copyrighted text [33]. LLM unlearning is fundamentally harder because knowledge is entangled across attention and MLP key-value memories, allowing multi-hop deductive reconstruction and evasion via paraphrasing [21], [34].

5. MATHEMATICAL AND THEORETICAL FOUNDATIONS

5.1 Empirical Risk Minimization and First-Order Stationarity

In statistical learning, empirical risk minimization on dataset D is formulated as:

$$\theta^* = \operatorname{argmin}_{\{\theta \in \Theta\}} \left[\frac{1}{|D|} \sum_{(x, y) \in D} \ell(f_{\theta}(x), y) + \frac{\lambda}{2} \|\theta\|_2^2 \right] \quad (13)$$

satisfying the first-order stationarity condition: $\nabla_{\theta} L(\theta^*; D) = \frac{1}{|D|} \sum \nabla_{\theta} \ell(f_{\theta^*}(x_i), y_i) + \lambda \theta^* = 0$.

5.2 Optimization Over the Retained Partition and Gradient Residuals

Partitioning $D = D_r \cup D_f$ decomposes total empirical loss as:

$$L(\theta; D) = \frac{N_r}{N} L(\theta; D_r) + \frac{N_f}{N} L(\theta; D_f) \quad (14)$$

The retrained parameters θ_{r^*} satisfy $\nabla_{\theta} L(\theta_{r^*}; D_r) = 0$. At the original parameter configuration θ^* , the retain gradient is non-zero:

$$\begin{aligned} \nabla_{\theta} L(\theta^*; D_r) &= - \frac{N_f}{N_r} \nabla_{\theta} L(\theta^*; D_f) \neq 0 \end{aligned} \quad (15)$$

Excising D_f creates an empirical gradient residual pulling parameters away from θ^* .

5.3 Taylor Series Approximations and the Newton Unlearning Step

Expanding $\nabla_{\theta} L(\theta_{r^*}; D_r)$ as a first-order Taylor series around θ^* yields $0 \approx \nabla_{\theta} L(\theta^*; D_r) + H_r (\theta_{r^*} - \theta^*)$, where $H_r = \nabla_{\theta}^2 L(\theta^*; D_r)$. Substituting the gradient residual yields the Newton unlearning step:

$$\begin{aligned} \theta_u &= \theta^* + \frac{1}{N_r} \sum_{z_j \in D_f} H_r^{-1} \nabla_{\theta} \ell(f_{\theta^*}(x_j), y_j) \end{aligned} \quad (16)$$

Under isotropic curvature ($H_r \approx \alpha I$), this reduces to first-order gradient ascent. Incorporating inverse Hessian H_r^{-1} scales updates along loss curvature axes, protecting retain capabilities [4], [29].

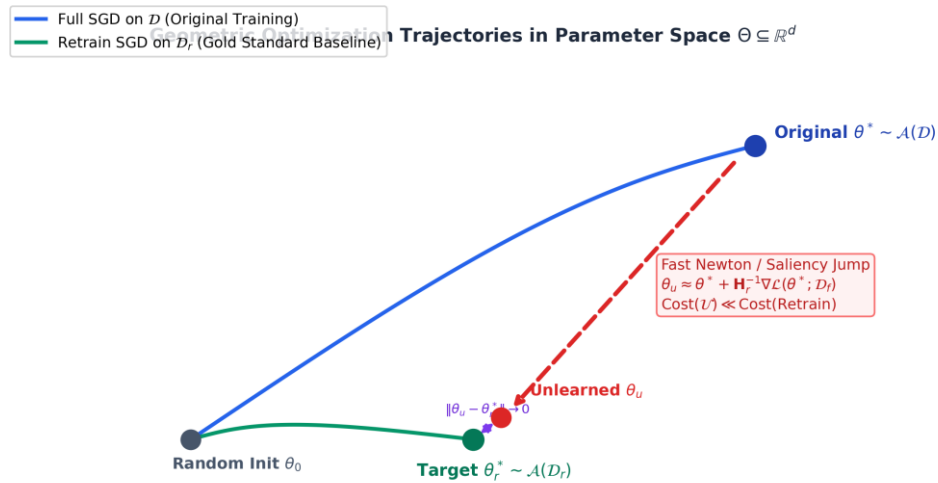


Fig. 5: Geometric optimization trajectories in parameter space. The Newton unlearning step provides a fast single-step approximation to the retrained target θ_{r^*} .

5.4 Parameter Distance Metrics versus Functional Divergence

Bounding parameter distance $\|\theta_u - \theta_{r^*}\|_2 \leq \xi$ is neither necessary nor sufficient: (1) in overparameterized networks, two close parameter points can produce wildly divergent predictions on outliers; and (2) two distant points can compute identical functions due to neuron permutation symmetries [38]. Rigorous frameworks bound distributional functional divergence:

$$\begin{aligned} W_1(P(f_{\{\theta_u\}}), P(f_{\{\theta_{r^*}\}})) &\leq \varepsilon \\ \text{or } D_{JS}(f_{\{\theta_u\}}(x) || f_{\{\theta_{r^*}\}}(x)) &\leq \varepsilon \quad (\forall x \in X) \end{aligned} \quad (17)$$

5.5 Certified (ε, δ) -Unlearning Formulations and Sensitivity Bounds

Certified unlearning adds Gaussian noise $b \sim N(0, \sigma^2 I)$ to the Newton update: $\theta_u = \theta^* + H_r^{-1} \nabla_{\theta} L(\theta^*; D_f) + b$, requiring noise scale $\sigma \geq \lceil \Delta_2 \sqrt{2 \ln(1.25 / \delta)} \rceil / \varepsilon$ with L_2 -sensitivity bound:

$$\Delta_2 \leq \lceil 2 L N_f \rceil / \lceil \mu N_r \rceil \quad (18)$$

where L is the gradient Lipschitz constant and μ is the strong convexity parameter. While this guarantees provable indistinguishability against arbitrary downstream membership probes [4], [7], [8], strict convexity assumptions do not hold in deep networks without local PL (Polyak-Łojasiewicz) approximations [9].

6. EVALUATION METRICS AND VERIFICATION OF UNLEARNING

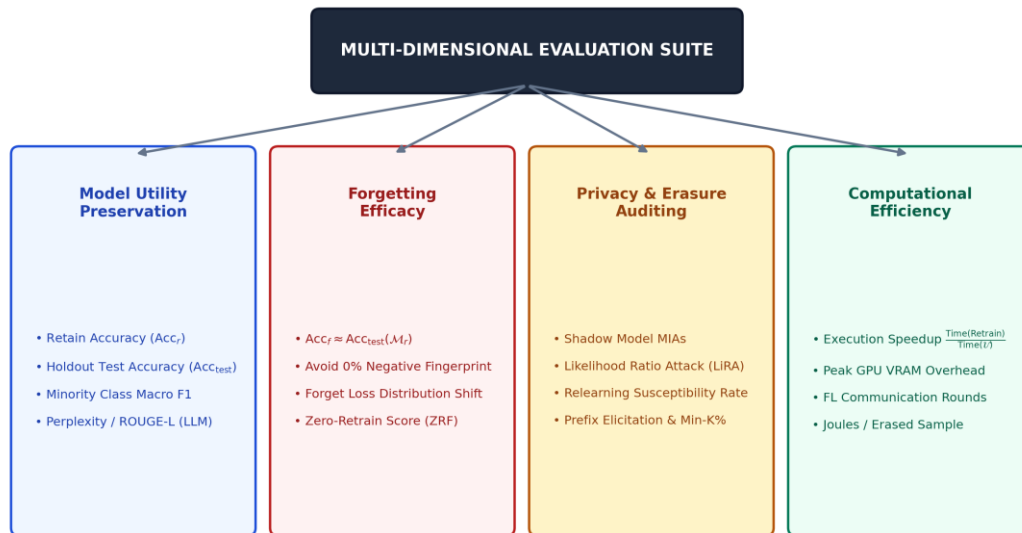


Fig. 6: Multi-dimensional evaluation suite for machine unlearning encompassing utility preservation, forgetting efficacy, privacy auditing, and computational efficiency.

6.1 Model Utility Preservation Metrics

Unlearning must preserve operational capabilities on authorized data: (1) Retain Accuracy (Acc_r), matching retrained reference $|Acc_r(Mu) - Acc_r(Mr)| \leq 0.5\%$; (2) Test Accuracy (Acc_{test}) on holdout sets; (3) Macro F1-Score across minority classes; and (4) Perplexity (PPL) in language models, avoiding degradation on standard corpora [21].

6.2 Forgetting Effectiveness Metrics

Forgetting metrics verify data excision: (1) Forget Accuracy (Acc_f), which must match random-guessing baseline or unseen test accuracy $Acc_f(Mu) \approx Acc_{test}(Mr)$ rather than 0% (which creates an identifiable negative fingerprint) [12], [28]; and (2) Forget Loss Distribution Shift, which should mirror unseen validation samples.

6.3 Membership Inference Attack (MIA) Auditing Frameworks

MIAs evaluate privacy empirically [18], [19]: (1) Shadow Model MIA, training classifiers on logit posteriors to predict membership [18]; and (2) Likelihood Ratio Attack (LiRA) [19], computing Gaussian log-ratios:

$$\Lambda(z) = \ln \left(\frac{p(f(z) | in)}{p(f(z) | out)} \right) = \ln p(f(z) | in) - \ln p(f(z) | out) \quad (19)$$

Effective unlearning drives MIA attack success down to random guessing (50.0%).

6.4 Reconstruction, Extraction, and Model Inversion Attacks

In generative architectures, forgetting is audited via: (1) Prefix Elicitation, checking if memorized text is output verbatim [17]; (2) Min-K% Probing, verifying token probability normalization on forget text [34]; and (3) Model Inversion, synthesizing class representations to confirm biometric feature eradication [28].

6.5 Comparative Divergence with Retrained Reference Models

Metrics evaluate structural alignment with retrained baseline Mr : (1) Zero-Retrain Forgetting (ZRF), measuring Jensen-Shannon divergence $ZRF = 1 - D_{JS}(f_{\theta_u}(Df) || f_{\theta_{r^*}}(Df))$ [10]; and (2) Centered Kernel Alignment (CKA) across internal representations [39].

6.6 Computational and Operational Efficiency Metrics

Assesses practical speedup: $\text{Speedup} = \text{Time}(\text{A}(\text{Dr})) / \text{Time}(\text{U})$, peak GPU VRAM allocation, and network communication volume in federated systems [15].

6.7 Verification Challenges: The Relearning Susceptibility Protocol

To distinguish genuine representation erasure from superficial decision boundary masking, researchers employ the Relearning Susceptibility Protocol [12], [21]: fine-tuning μ for 1–3 epochs on retain data. Masked models rapidly recover forget capabilities (>70% in one epoch), whereas genuinely unlearned models recover slowly, matching clean baseline learning dynamics.

Table 2: Multi-Dimensional Evaluation Metrics and Verification Protocols

Evaluation Dimension	Primary Metric(s)	Target Value	Failure Mode Detected
Model Utility	Retain Accuracy (Acc_r), Test Accuracy (Acc_test)	$\approx \text{Acc}(\text{Mr})$	Catastrophic forgetting; collateral retain damage
Generative Utility	Perplexity (PPL), ROUGE-L, BLEU	$\leq 1.05 \times \text{PPL}(\text{Mr})$	Language degradation; repetitive loops
Forgetting Efficacy	Forget Accuracy (Acc_f), Cross-Entropy on Df	$\approx \text{Acc_test}(\text{Mr})$	Under-unlearning or over-unlearning (0% fingerprint)
Privacy Auditing	Membership Inference Attack Success (MIASR)	$\approx 50.0\%$ (Random baseline)	Residual latent memorization; privacy leaks
Rigorous Privacy	Likelihood Ratio Attack (LiRA) Log-Ratio [19]	Matches holdout distribution	Confidence score distribution anomaly
Representation Erasure	Relearning Susceptibility Rate [12]	Gradual recovery matching Mr	Decision boundary masking; latent weight persistence
Computational Gain	Execution Wall-Clock Speedup Factor	$>> 5.0 \times (10 \times -100 \times)$	Inefficiency approaching full retraining overhead

7. COMPARATIVE ANALYSIS OF MACHINE UNLEARNING METHODS

Table 3: Comparison of Major Machine Unlearning Techniques

Technique	Accuracy (Acc_r)	Speedup	Privacy Guarantee	Scalability	Main Limitation	Technical
Exact Retraining	Optimal (Baseline)	1.0×	Flawless: Zero statistical exposure	Very Poor: Infeasible for LLMs	Quadratic cost explosion under streaming requests	
SISA Sharding [2]	Near-Optimal	S×	Exact: Sub-shard never saw Df	Moderate: Scales with shard count	S-fold inference latency; lower data diversity	
Gradient Ascent [5]	Poor to Moderate	20×–100×	None: Heuristic only; vulnerable to LiRA	High: Readily applied to large nets	Decision boundary masking; instability	
Influence Functions [29]	High (Convex only)	Fast (1-step)	Asymptotic / Certified (Convex)	Very Poor: Intractable for $d > 10^7$	Non-convexity violations; negative eigenvalues	
Distillation (SCRUB) [12]	Very High (<0.2% drop)	4×–8×	Empirical: Strong resistance to shadow MIAs	High: Scales to ResNets, ViTs, LLMs	Requires teacher caching; slower than pruning	
Saliency (SalUn/SSD) [13]	High (Freezes weights)	8×–25×	Empirical: Targeted weight modification	Very High: Readily computed on large nets	Threshold sensitivity; traces in frozen layers	
Certified Noise [4], [8]	Moderate to Low	Fast (1-step)	Provable: Cryptographic (ϵ, δ) bound	Poor to Moderate: Noise scales with d	Severe utility collapse in overparameterized nets	
FedEraser [15]	High	3×–5×	Empirical to Certified (With DP noise)	High: Designed for edge setups	Trajectory drift over long aggregation histories	

Table 4: Machine Unlearning vs. Related Concepts

Feature	Machine Unlearning	Retraining	Data Deletion	Differential Privacy [20]	Model Editing [22]
---------	--------------------	------------	---------------	---------------------------	--------------------

Feature	Machine Unlearning	Retraining	Data Deletion	Differential Privacy [20]	Model Editing [22]
Objective	Revoke influence of Df from weights	Produce model with zero exposure	Purge rows from physical disk	Bound privacy leakage of all data	Alter specific factual associations
Data Removal	Conceptual & functional in weights	Complete physical & statistical exclusion	Physical row deletion from disk	Data retained; sensitivity bounded	Data retained; associations altered
Model Impact	Post-hoc adjustment of parameters θ	Complete re-optimization from scratch	Zero change: Model untouched	Clipped gradients & noise in training	Direct update to MLP projections
Privacy Bound	Exact, approximate, or (ϵ, δ) bound	Flawless: $I(\theta_{r^*}; D_f D_r) = 0$	Zero privacy protection for model	Provable (ϵ, δ) bound on all data	No formal privacy guarantee
Compute Cost	$\ll O(N)$ passes; high speedup	$O(N \cdot E)$ epochs; prohibitive	$O(1)$ command; compute	RDBMS zero training overhead	Ultra-fast: Few seconds per edit
Applications	GDPR Art. 17, CCPA, toxic data	Small academic benchmark datasets	Database maintenance & cleanup	Enterprise aggregate analytics	Correcting outdated facts / hallucinations

7.1 Comparative Synthesis across Performance, Privacy, and Scalability Axes

Tables 3 and 4 delineate clear operational boundaries: (1) Exact vs. Approximate, where exact methods (Retraining, SISA) eliminate legal ambiguity at high compute/latency costs, while approximate heuristics (SalUn, SCRUB) deliver large speedups (5×–25×) requiring empirical MIA verification; (2) Unlearning vs. Differential Privacy, where DP-SGD introduces permanent utility loss across all data during training, while unlearning acts reactively, preserving full accuracy until deletions occur; and (3) Unlearning vs. Model Editing, where editing patches factual queries without dissolving latent representations, failing privacy auditing.

8. APPLICATIONS OF MACHINE UNLEARNING

8.1 Healthcare and Clinical Decision Systems

Clinical diagnostic models trained on electronic health records (EHR) and radiological scans contain private patient data. When trial participants withdraw consent or samples are contaminated, unlearning purges the cohort without taking diagnostic infrastructure offline [35].

8.2 Financial Services and Credit Systems

In automated credit underwriting and fraud detection [36], Fair Credit Reporting acts and GDPR empower users to dispute erroneous records. Unlearning enables financial institutions to purge biased or disputed customer cohorts, mitigating algorithmic discrimination.

8.3 Recommendation Systems and Consumer Personalization

E-commerce and streaming models (e.g., DLRM) continuously train on user click-streams. When users close accounts or request a "clean slate" preference reset, unlearning recalibrates user embedding spaces without retraining the entire recommendation pipeline [2].

8.4 Social Media and User-Generated Platforms

When digital creators withdraw licensing rights or users delete accounts, unlearning excises personal image portfolios and biometric identities from content-tagging and face-recognition classifiers [33].

8.5 Decentralized Edge and Federated Learning

In federated edge deployments (e.g., mobile keyboard prediction), unlearning protocols (FedEraser, FedRecovery) purge departing clients' historical parameter updates without requiring costly global retraining rounds across thousands of edge devices [15], [16].

8.6 Computer Vision, Biometrics, and Surveillance Systems

When security personnel depart sensitive facilities, class-unlearning and identity-pruning algorithms purge facial identity vectors while preserving recognition capabilities for authorized staff [5], [30].

8.7 Generative Foundation Models and Large Language Models

In foundation models, unlearning mitigates major legal and safety risks [21], [24]: (1) scrubbing copyrighted literature to settle IP infringement suits [33]; (2) neutralizing dangerous CBRN instructions (WMDP / RMU) [24]; and (3) removing private PII harvested during uncurated web scrapes [17], [21].

Table 5: Real-World Applications, Threat Profiles, and Operational Requirements

Application Domain	Forget Target (Df)	Primary Threat Vector	Permissible Latency	Required Standard	Privacy	Recommended Strategy
Clinical Healthcare	Patient scans, EHR histories	Re-identification via LiRA probes [19]	Hours to Days	Certified / Exact bound	Exact ((ε, δ)	SISA Partitioning / Certified Perturbation [2], [4]
Financial Services	Disputed credit records, biased cohorts	Reconstruction of financial profiles [18]	Minutes to Hours	Statutory Compliance (FCRA / GDPR)	Compliance	Influence Functions / Relational Distillation [12], [29]
E-Commerce & RecSys	Deactivated accounts, transient logs	Behavioral profile extraction [35]	Real-time Minutes	Empirical Shift	Boundary	Saliency Masking / Synaptic Dampening [13], [14]
Decentralized (FL)	Edge Departing client's historical updates	Model inversion of client updates [15]	Communication rounds	Client-level Anonymity		FedEraser / FedRecovery [15], [16]
Computer Vision	Biometric faces, copyrighted artwork	Class inversion; facial recognition probes	Minutes to Hours	Complete Erasure	Feature	Saliency-Masked Unlearning (SalUn) [14]
Foundation LLMs	Models / Toxic text, CBRN instructions, books	Prefix elicitation; Min-K% extraction [17], [34]	Hours (Alignment tuning)	Functional & Deductive Eradication		Representation Misdirection (RMU) / DPO [24], [25]

9. CHALLENGES AND LIMITATIONS

9.1 Computational Complexity and Second-Order Overhead

Exact second-order unlearning requires calculating and inverting Hessian curvature matrices [29], [38]. In models with billions of parameters, computing even diagonal Fisher approximations exceeds GPU memory, requiring costly distributed approximations.

9.2 The Fundamental Verification Dilemma

Third-party auditors cannot easily verify whether private data has been purged. In black-box settings, unlearning routines can alter decision boundaries to suppress outputs on forget queries while leaving latent representations intact in deeper layers [38].

9.3 Catastrophic Forgetting and Collateral Utility Damage

Modifying parameters to excise target memories often damages adjacent representations. In fine-grained classification, unlearning one class frequently degrades accuracy across related classes sharing semantic features [12], [14].

9.4 The Pareto Frontier: Privacy Guarantees versus Retain Utility

Aggressive unlearning (large gradient steps or high noise) guarantees complete data eradication but collapses retain accuracy and spikes language model perplexity. Conservative updates preserve utility but leave vulnerable residual traces [4], [12].

9.5 Sequential and Streaming Unlearning Degradation

In production systems, deletion requests arrive sequentially: $M_0 \rightarrow _U M_1 \dots \rightarrow _U M_K$. Repeatedly applying approximate heuristics causes approximation errors to compound, resulting in severe weight drift and model collapse within 20–50 cycles [38].

9.6 Scalability Bottlenecks in Foundation Architectures

In models trained on trillions of tokens, knowledge is deeply entangled across self-attention heads and feed-forward layers. Editing parameters to remove documents often breaks syntactic coherence and causes repetitive loops [21], [22].

9.7 Fragmentation and Lack of Standardized Benchmarks

Historically, inconsistent forget ratios, evaluation metrics, and architectures across research groups have obscured whether empirical gains represented genuine algorithmic advances or hyperparameter artifacts [28].

9.8 Large Language Model-Specific Structural Complications

In LLMs, unlearning faces: (1) multi-hop relational deduction, where unlearned facts are deduced from surrounding context; (2) in-context re-emergence under prompt framing; and (3) cross-lingual transfer failures [21], [34].

10. SECURITY AND PRIVACY CONSIDERATIONS

10.1 The Privacy Inversion Paradox and Differential Parameter Leakage

Executing an unlearning request paradoxically confirms that the target data was present in training. Observing drops in model confidence or weight difference vectors $\Delta\theta = \theta_{\text{post}} - \theta_{\text{pre}} \approx \eta \nabla_{\theta} L(\theta; D_f)$ allows attackers to verify training set membership or reconstruct private forget samples via gradient inversion attacks [26].

10.2 Malicious Deletion Requests and Adversarial Unlearning

Adversaries can exploit unlearning APIs through: (1) targeted algorithmic sabotage, deleting critical edge cases (e.g., pedestrian imagery in autonomous perception); (2) resource exhaustion Denial-of-Service (DoS), submitting requests for critical support vectors to degrade utility for all users; and (3) poisoned unlearning triggers, crafting training points that destabilize parameters when unlearned [27].

10.3 Architectural Defense-in-Depth for Secure Unlearning Operations

Mitigating these risks requires: (1) cryptographic identity authentication for all deletion requests; (2) batched, randomized execution schedules paired with differential privacy noise injection to obfuscate parameter trajectories [4], [8]; and (3) immutable, zero-knowledge compliance audit logs [35].

11. RECENT RESEARCH TRENDS (2020–2026)

11.1 Transition Toward Certified and Provable Privacy Guarantees

Recent scholarship has transitioned from empirical heuristics toward provable bounds, leveraging local Polyak-Łojasiewicz (PL) geometry and algorithmic stability to certify (ϵ, δ) bounds on non-convex landscapes [8], [9].

11.2 Sparse Sub-Network Localization and Saliency Pruning

Recognizing that global weight updates cause collateral damage, methods such as SSD [13] and SalUn [14] demonstrate that confining updates to the top 1%–5% salient weights achieves near-gold-standard unlearning with minimal retain loss and order-of-magnitude speedups.

11.3 Standardized Foundation Model Benchmarks (TOFU, WMDP, MUSE)

Recent benchmarks provide standardized evaluation: TOFU [21] for synthetic author unlearning in LLaMA; WMDP and Representation Misdirection Unlearning (RMU) [24] for neutralizing CBRN weapon knowledge; and Preference Optimization Unlearning (DPO) [25] for steering generative distributions.

12. RESEARCH GAPS AND FUTURE DIRECTIONS

Despite recent progress, key research gaps remain unresolved across thirteen critical research directions:

- **1. Certified Unlearning for Foundation Models:** Global DP noise scales poorly with parameter dimension d . Direction: Formulate subspace-restricted certified bounds over low-rank parameter manifolds (e.g., LoRA).
- **2. Multi-Hop Relational Erasure:** Erasing isolated facts leaves deductive pathways intact in LLMs. Direction: Integrate causal mediation analysis and knowledge graphs to sever complete relational dependency chains.
- **3. Multi-Modal Benchmarks:** Vision-language models exhibit cross-modal leakage. Direction: Build unified benchmarks evaluating joint cross-modal representation erasure.
- **4. Sequential Streaming Unlearning:** Repeated approximate unlearning causes cumulative error drift. Direction: Develop drift-correcting algorithms with anchor regularization and recursive calibration.
- **5. Zero-Knowledge Verification (ZK-PoU):** Verifying unlearning currently risks secondary data leaks. Direction: Construct Zero-Knowledge Proofs of Unlearning to verify compliance without revealing weights or data.
- **6. Mitigating Relearning Susceptibility:** Models frequently mask decision boundaries rather than erasing features. Direction: Incorporate adversarial representation scattering to dissolve latent feature manifolds permanently.
- **7. Adversarially Robust Unlearning:** APIs are vulnerable to DoS and poisoning attacks. Direction: Implement influence-based anomaly detection to filter malicious deletion requests.
- **8. Thermodynamic & Green Unlearning:** Lifecycle energy consumption is rarely audited. Direction: Establish standardized thermodynamic benchmarks measuring Joules-per-erased-sample.

- **9. Automated Adaptive Tuning:** Heuristics are hypersensitive to hyperparameters. Direction: Develop self-terminating unlearning routines based on internal representation divergence monitoring.
- **10. Statutory and Regulatory Alignment:** Disconnect between legal definitions and statistical metrics. Direction: Interdisciplinary frameworks formalizing statistical unlearning thresholds under GDPR/CCPA.
- **11. Asynchronous Non-IID Federated Unlearning:** Edge networks face client dropouts and data skew. Direction: Design asynchronous federated operators reconstructing clean global states despite missing update logs.
- **12. Black-Box API-Boundary Unlearning:** Commercial models (e.g., GPT-4) permit no weight access. Direction: Formulate inference-time representation steering and output-filtering caches.
- **13. Unlearning in Reinforcement Learning:** Sequential policies embed experiences in non-stationary trajectories. Direction: Formulate policy-gradient unlearning algorithms excising unsafe experiences without policy retraining.

13. CRITICAL DISCUSSION

13.1 Which Machine Unlearning Techniques Appear Most Practical for Real-World Deployment?

For small-to-medium networks requiring exact compliance, SISA data partitioning is the most defensible choice. For deep vision models, saliency masking (SalUn) and relational distillation (SCRUB) offer the best empirical Pareto trade-off (5×–10× speedup with minimal utility loss). For LLMs, representation steering (RMU) and preference optimization (DPO) are currently the only scalable strategies that maintain linguistic coherence.

13.2 When is Exact Unlearning Preferable, and When are Approximate Methods Acceptable?

Exact unlearning is mandatory in high-liability, audited domains (litigation, classified defense data, biometric databases). Approximate unlearning is acceptable in high-throughput consumer services (recommendation feeds, social media tagging) where requests arrive continuously and massive compute reductions outweigh minor theoretical leakage risks.

13.3 What are the Major Trade-offs Between Privacy and Model Utility?

Aggressive unlearning completely eliminates membership inference leakage but damages retain accuracy and inflates language perplexity. Conservative updates preserve utility but leave residual latent traces. Hyperparameters must be tuned to match domain-specific threat models.

13.4 Are Current Unlearning Evaluation Methods Sufficient?

No. Simple classification accuracy on the forget set is fundamentally flawed because models can mask decision boundaries while preserving latent representations. Robust verification requires multi-dimensional suites combining shadow MIAs, LiRA log-ratios, loss distribution shifts, and relearning susceptibility protocols.

13.5 Can Current Techniques Scale to Foundation Models?

Exact and second-order methods cannot scale to multi-billion-parameter foundation models due to memory and compute bounds. Only first-order localized approaches, parameter-efficient fine-tuning (LoRA), and representation steering demonstrate practical feasibility at web scale.

13.6 Is Machine Unlearning a Complete Solution to the Right to be Forgotten?

No. Unlearning addresses only model parameter weights. True legal compliance under GDPR Article 17 requires end-to-end data governance: purging raw storage databases, clearing data lake feature stores, updating downstream ensembles, and maintaining immutable audit trails.

13.7 What Theoretical Guarantees are Still Missing?

The most critical missing theoretical guarantee is dimension-independent certified privacy bounds for non-convex neural representations. Developing certified indistinguishability without injecting noise that collapses overparameterized utility remains the central open theoretical challenge.

14. CONCLUSION

Machine Unlearning has rapidly evolved into a cornerstone discipline for trustworthy artificial intelligence. Driven by global privacy regulations such as GDPR and CCPA, as well as the imperative to scrub toxic, copyrighted, and memorized personal data from massive foundation models, unlearning bridges the divide between statutory rights and statistical learning.

This paper has presented an academically rigorous, comprehensive survey of machine unlearning. We mathematically formalized the unlearning problem under Empirical Risk Minimization, analyzed exact, approximate, and certified paradigms, and presented an exhaustive taxonomy spanning SISA sharding, gradient inversion, influence functions, distillation, parameter saliency, certified perturbation, and federated unlearning. Furthermore, we examined foundation model unlearning, where entangled representations and multi-hop reasoning present unique technical hurdles.

Our analysis demonstrates that evaluating unlearning purely through classification accuracy is fundamentally misleading. Robust evaluation requires multi-dimensional auditing suites incorporating shadow-model MIAs, likelihood ratio tests (LiRA), loss-distribution shifts, and relearning susceptibility protocols. While approximate methods achieve dramatic computational speedups over full retraining, addressing sequential drift, catastrophic utility collapse, and unlearning-induced privacy leakage through subspace-restricted certified bounds, auditable zero-knowledge proofs, and cross-modal erasure will define the next generation of trustworthy AI systems.

REFERENCES

- [1] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in 2015 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2015, pp. 463-480.
- [2] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in 2021 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2021, pp. 141-159.
- [3] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou, "Making AI forget you: Data deletion in machine learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, Vancouver, Canada, 2019, pp. 5125-5136.
- [4] C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten, "Certified data removal from machine learning models," in Proceedings of the 37th International Conference on Machine Learning (ICML), Vienna, Austria, 2020, pp. 3832-3842.
- [5] A. Golatkar, A. Achille, and S. Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9304-9312.
- [6] A. Golatkar, A. Achille, and S. Soatto, "Forgetting outside the box: Scrubbing deep networks of information access from input-output observations," in Computer Vision - ECCV 2020, Glasgow, UK, 2020, pp. 383-398.
- [7] A. Sekhari, J. Acharya, G. Kamath, and S. U. Wu, "Remember what you want to forget: Algorithms for machine unlearning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, Virtual, 2021, pp. 18075-18086.
- [8] S. Neel, A. Roth, and S. Sharifi-Malvajerdi, "Descent-to-delete: Gradient-based methods for machine unlearning on convex summands," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, Virtual, 2021, pp. 13573-13584.
- [9] E. Ullah, T. Mhamdi, R. Caruana, and R. Khordon, "Machine unlearning via algorithmic stability," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, Virtual, 2021, pp. 8226-8237.
- [10] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. V. Kankanhalli, "Zero-shot machine unlearning," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2345-2354, 2023.
- [11] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Can bad teaching induce forgetting? Unlearning in deep networks using an incompetent teacher," in Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 2023, pp. 3584-3592.
- [12] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou, "Towards unbounded machine unlearning," in Proceedings of the 11th International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023.
- [13] J. Foster, S. Schoelkopf, and A. Brintrup, "Fast machine unlearning without retraining through selective synaptic dampening," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 11, Vancouver, Canada, 2024, pp. 12043-12051.
- [14] C. Fan, J. Liu, Y. Zhang, D. Wei, E. Wong, and S. Liu, "SalUn: Empowering machine unlearning via gradient-based weight saliency," in Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 2024.
- [15] G. Liu, X. Ma, T. Yang, C. Wang, and J. Liu, "FedEraser: Enabling efficient client-level data removal on federated learning systems," in 2021 IEEE 29th International Conference on Network Protocols (ICNP), Dallas, TX, USA, 2021, pp. 1-10.
- [16] L. Wang, T. Dong, and F. Liu, "FedRecovery: Differentially private federated unlearning with active client selection," in Proceedings of the 32nd USENIX Security Symposium (USENIX Security), Anaheim, CA, USA, 2023, pp. 3121-3138.
- [17] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in Proceedings of the 30th USENIX Security Symposium (USENIX Security), Virtual, 2021, pp. 2633-2650.
- [18] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017, pp. 3-18.
- [19] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in 2022 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2022, pp. 1897-1914.
- [20] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Theory of Cryptography Conference (TCC), New York, NY, USA, 2006, pp. 265-284.
- [21] P. Maini, Z. Feng, A. Schwarzschild, Z. C. Lipton, and J. Z. Kolter, "TOFU: A task of fictitious unlearning for large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, Vancouver, Canada, 2024.
- [22] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, New Orleans, LA, USA, 2022, pp. 17359-17372.
- [23] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau, "Mass-editing memory in a transformer," in Proceedings of the 11th International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023.
- [24] A. Zou, L. Phan, H. Wang, D. Duenas, N. Long, N. Li, S. Casper, M. Mazeika, and D. Hendrycks, "The WMDP benchmark: Measuring and mitigating bio and cyber knowledge in LLMs," in Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 2024.

- [25] Z. Yao, X. Chen, and Y. Mao, "Large language model unlearning via preference optimization," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand, 2024, pp. 8412-8425.
- [26] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning meets membership inference attacks," in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS), Virtual, 2021, pp. 756-774.
- [27] N. G. Marchant, B. I. Rubinstein, and S. Alfeld, "Hard to forget: Poisoning attacks on certified machine unlearning," in Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS), Los Angeles, CA, USA, 2022, pp. 783-796.
- [28] B. Di, Y. Guo, S. Zhang, and X. Liu, "Towards realistic evaluation of machine unlearning," in Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 2024.
- [29] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 2017, pp. 1885-1894.
- [30] A. Warnecke, L. Pirch, C. Wressnegger, and K. Rieck, "Machine unlearning of features and labels," in 2023 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2023, pp. 227-244.
- [31] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. V. Kankanhalli, "Fast yet effective machine unlearning," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 7812-7824, 2024.
- [32] J. Jia, J. Liu, and B. Li, "Model sparsity can simplify machine unlearning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, New Orleans, LA, USA, 2023.
- [33] R. Eldan and M. Russinovich, "Who's Harry Potter? Approximate unlearning in LLMs," arXiv preprint arXiv:2310.02238, 2023.
- [34] W. Shi, C. Ajith, M. Sakurovs, Y. Zhang, B. Wang, and T. Hashimoto, "Detecting pretraining data from large language models," in Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 2024.
- [35] H. Zhang, T. Nakamura, T. Isohara, and K. Sakurai, "A survey on machine unlearning: Taxonomy, metrics, and future directions," ACM Computing Surveys, vol. 56, no. 4, pp. 1-38, 2023.
- [36] J. Xu, Z. Wu, C. Wang, and X. Jia, "Machine unlearning: A survey," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 36, no. 12, pp. 7512-7531, 2024.
- [37] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning," ACM Computing Surveys, vol. 55, no. 11, pp. 1-36, 2022.
- [38] A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot, "Unrolling SGD: Understanding factors influencing machine unlearning," in 2022 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2022, pp. 1915-1932.
- [39] S. Goel, A. Bansal, S. Sanjabi, B. He, and R. Salakhutdinov, "Evaluating representation removal for machine unlearning," in Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 2024.
- [40] X. Li, J. Tan, and Z. Chen, "Making large language models forget with few demonstration examples," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, New Orleans, LA, USA, 2023.
- [41] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," arXiv preprint arXiv:2001.08361, 2020.
- [42] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., "Training compute-optimal large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 30016-30030.
- [43] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," arXiv preprint arXiv:2104.10350, 2021.