Towards Benchmarking Privacy Vulnerabilities in Selective Forgetting with Large Language Models

Wei Qian[*1], **Chenxu Zhao**[*1], **Yangyi Li**[1], **Mengdi Huai**[1]

[1]Iowa State University
{wqi, cxzhao, liyangyi, mdhuai}@iastate.edu

Abstract

The rapid advancements in artificial intelligence (AI) have primarily focused on the process of learning from data to acquire knowledgeable learning systems. As these systems are increasingly deployed in critical areas, ensuring their privacy and alignment with human values is paramount. Recently, *selective forgetting* (also known as *machine unlearning*) has shown promise for privacy and data removal tasks, and has emerged as a transformative paradigm shift in the field of AI. It refers to the ability of a model to selectively erase the influence of previously seen data, which is especially important for compliance with modern data protection regulations and for aligning models with human values. Despite its promise, selective forgetting raises significant privacy concerns, especially when the data involved come from sensitive domains. While new unlearning-induced privacy attacks are continuously proposed, each is shown to outperform its predecessors using different experimental settings, which can lead to overly optimistic and potentially unfair assessments that may disproportionately favor one particular attack over the others. In this work, we present the first comprehensive benchmark for evaluating privacy vulnerabilities in selective forgetting. We extensively investigate privacy vulnerabilities of machine unlearning techniques and benchmark privacy leakage across a wide range of victim data, state-of-the-art unlearning privacy attacks, unlearning methods, and model architectures. We systematically evaluate and identify critical factors related to unlearning-induced privacy leakage. With our novel insights, we aim to provide a standardized tool for practitioners seeking to deploy customized unlearning applications with faithful privacy assessments.

Introduction

In recent years, artificial intelligence (AI) has revolutionized nearly every aspect of modern life. In an era of AI, the primary challenge is enabling models to acquire broad knowledge effectively. However, the training datasets employed in training these models often contain sensitive information encompassing private and copyrighted content (Bu et al. 2024; Mueller et al. 2024; Chu, Song, and Yang 2024; Wei et al. 2024). This situation raises significant risks of sensitive data leakage, directly conflicting with the growing legislative emphasis on the "right to be forgotten" (Bukaty 2019; Regulation 2018). Instances such as the proliferation of copyright infringement cases post the release of models (Rombach et al. 2022), and The New York Times's lawsuit against OpenAI for content leakage (Hadero and Bauder 2023), underscore the urgency of addressing these issues.

In response to these challenges, *selective forgetting* (also referred to as *machine unlearning*) (Li et al. 2025; Zhang et al. 2024; Qian et al. 2023; Zhao et al. 2023; Qian et al. 2022; Bourtoule et al. 2021) has emerged as a promising solution. Selective forgetting aims to compel models to forget sensitive information without retraining, thereby eliminating the risk of content leakage. In contrast, retraining models from scratch to accommodate deletions is impractical due to the extensive computational resources required. Machine unlearning aims to remove the influence of *the requested unlearning data* from a pre-trained model, producing an unlearned model that approximates one retrained from scratch using only *the retain data* (i.e., the original training data excluding the unlearning set). Recent work also leverages conformal prediction (Li and Huai 2025) to quantify forgetting uncertainty, leading to more rigorous unlearning (Alkhatib and Tay 2025). Machine unlearning not only aids in meeting regulatory requirements but also enhances learning systems' privacy protection by sensitive data attacks.

However, adopting machine unlearning techniques may not always provide the anticipated privacy protections, and could even introduce new privacy vulnerabilities. First, for the requested unlearning data, machine unlearning naturally generates two versions of machine learning models, namely the original model and the unlearned model, which differ due to the deletion of the unlearning data. This discrepancy can inadvertently leak information about the unlearned data. Additionally, the unlearned model alone may still retain residual privacy risks related to the requested unlearning data due to incomplete forgetting. Moreover, the privacy of the unlearning data may be further compromised when the unlearned model is subjected to future deployment scenarios such as fine-tuning, which may reactivate or amplify memorized knowledge. Even worse, beyond the privacy risks of the unlearning data, selective forgetting may also influence the retain data, potentially altering their privacy exposure through model shifts or malicious fingerprints. These concerns highlight that selective forgetting may introduce new

[*]These authors contributed equally.

privacy attack surfaces that adversaries can exploit, potentially undermining the guarantees associated with unlearning requests or compromising the privacy of other data.

Currently, many works have been proposed to investigate privacy risks stemming from selective forgetting (Hu et al. 2025; Łucki et al. 2025; Zhang et al. 2025; Yuan et al. 2025; Wang et al. 2025a; Hu et al. 2024; Carlini et al. 2022b; Lu et al. 2022a; Chen et al. 2021). For example, (Hu et al. 2024) examines data reconstruction attacks (DRAs) on unlearning data by leveraging the discrepancy between the pre-trained and unlearned models. However, there still lacks a structured understanding of the empirical privacy risks of machine unlearning techniques. Without a clear understanding of the practical risks, practitioners are left with little guidance on how to safely and privately apply machine unlearning techniques in privacy-sensitive settings. Additionally, existing unlearning-induced privacy attacks are typically evaluated under disparate experimental settings, with varying experimental settings. As a result, each of them is often shown to outperform prior methods under its own tailored conditions, leading to overly optimistic and potentially inconsistent evaluations that may unfairly favor certain attacks. Consequently, an in-depth investigation into the effectiveness of unlearning-induced privacy vulnerabilities in a standard and reproducible experimental setting is missing.

To address these limitations, we in this paper introduce the first comprehensive benchmark **PrivUB**, i.e., **Priv**acy Vulnerabilities in Machine **U**nlearning **B**enchmark. This work makes four major contributions:

(1) We present the first benchmark that systematically evaluates existing privacy vulnerabilities introduced by machine unlearning. Our benchmark emphasizes the importance of aligning privacy guarantees with human intent, highlighting gaps between technical implementations and user expectations. Our benchmark reveals fundamental challenges in unlearning and provides a critical foundation for understanding its implications in the context of emerging data protection regulations and broader challenges in AI alignment.

(2) We instantiate a structured taxonomy of privacy vulnerabilities in machine unlearning by implementing representative attacks across key dimensions, including privacy vulnerability type, victim data type, victim model type, and attacking tool. Each is grounded with a specific threat model.

(3) We evaluate existing defense methods targeting privacy risks in machine unlearning, analyzing their effectiveness in a structured manner across different types of privacy vulnerabilities, victim data, victim model, and attacking tool.

(4) Through extensive empirical studies, we conduct a comprehensive evaluation covering 21 unlearning-induced privacy attack and defense methods in machine unlearning, 11 real-world datasets, 10 mainstream models, 10 popular unlearning techniques, and 10 task-specific evaluation metrics.

We present a thorough analysis of the above evaluations from different perspectives to examine privacy vulnerabilities introduced by selective forgetting. Our key findings include: (1) Combining multiple attacking tools (including perturbing unlearned model and perturbing unlearned data) can improve attack effectiveness. (2) The attacking tools of perturbing unlearned data designed for knowledge leakage attacks can be utilized to further enhance the performance of membership inference attacks (MIAs). (3) The privacy risks caused by the fine-tuning method are more severe than those caused by the model quantization method. (4) Existing privacy attacks, with proper adaptation, can be successfully generalized across model types. Notably, we find that attacks originally developed for deep learning models can be applied to large language models (LLMs), and vice versa, while maintaining strong performance. (5) Existing defenses against privacy vulnerabilities generally lack robustness. In particular, some defenses are highly sensitive to the number of attack samples, leading to inconsistent protection.

Related Work

The rapid development of machine learning models has significantly benefited various applications. However, their increasing deployment has raised serious privacy concerns, particularly in sensitive domains such as healthcare and finance. Notably, models often unintentionally memorize their training data, going beyond merely learning the general patterns within the data. This behavior makes models vulnerable to various privacy attacks, including membership inference attacks (Zhao et al. 2025; Carlini et al. 2022a; Chen et al. 2021), data reconstruction attacks (Hu et al. 2024; Du et al. 2024; Yuan et al. 2023), and knowledge leakage attacks (Hu et al. 2025; Łucki et al. 2025; Yuan et al. 2025).

Currently, many privacy benchmarks have been proposed to investigate privacy risks associated with machine learning models (Niu et al. 2025; Wen et al. 2025; Chen et al. 2025; Zhu et al. 2024; Li et al. 2023; Song and Mittal 2021). For example, (Niu et al. 2025) presents a systematic comparison of various membership inference attacks using carefully designed evaluation scenarios. Rigorous privacy evaluation is essential for identifying vulnerabilities in models, and developing a comprehensive understanding of existing research gaps and potential mitigation strategies, thereby promoting alignment with privacy principles and human values. In this work, we aim to benchmark privacy vulnerabilities in selective forgetting. This is the first benchmark to systematically study the unlearning-induced privacy attacks and defenses.

Benchmark Framework

Setup of Privacy Evaluation

Let $f(\cdot;\theta)$ denote the pre-trained model, where $\theta \in \Theta$ denotes the model parameters. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote a dataset of n samples drawn from an underlying distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. Note that *during the training phase*, a training algorithm T maps the training dataset $\mathcal{D}$ to parameters θ, yielding the pre-trained model. During *the unlearning phase*, the pre-trained model is updated by an unlearning algorithm U, which aims to remove the influence of the requested unlearning data, which is typically a subset of the training data $\mathcal{D}$. During *the deployment phase*, the model can be further modified via a fine-tuning procedure F on task-specific data, or a quantization operator Q, which compresses the parameters θ for efficient inference. Below, we elaborate the unlearning phase and the deployment.

Privacy vulnerability type	Victim data type	Victim model type	Attacking tool	Paper	Threat model			Architecture
					Model access information	A	V	
Membership inference	Unlearning data	Pre-trained and unlearned models	Pre-trained and unlearned model discrepancy	(Chen et al. 2021)	Posterior/ Top-k posterior/ Label-only	Yes	Yes	Deep learning
				(Lu et al. 2022a)	Label-only	Yes	Yes	Deep learning
				(Lu et al. 2022b)	Label-only	Yes	Yes	Deep learning
				(Du et al. 2024)	Loss	No	No	LLM
	Retain data	Unlearned model	Malicious unlearning subset	(Carlini et al. 2022b)	Model weights	No	Yes	Deep learning
				(Gu, He, and Chen 2024)	Model weights	No	Yes	Deep learning
Data reconstruction	Unlearning data	Pre-trained and unlearned models	Pre-trained and unlearned model discrepancy	(Hu et al. 2024)	Model weights	No	Yes	Deep learning
				(Du et al. 2024)	Model weights	No	Yes	LLM
				(Wang et al. 2025a)	Posterior	Yes	Yes	Deep learning
Knowledge leakage	Unlearning data	Fine-tuned model	Perturbing unlearned model	(Doshi and Stickland 2024)	Model weights	No	Yes	LLM
				(Hu et al. 2025)	Model weights	No	Yes	LLM
				(Zhang et al. 2025)	Model weights	No	Yes	LLM
				(Łucki et al. 2025)	Model weights	No	Yes	LLM
		Unlearned model	Perturbing unlearned data	(Doshi and Stickland 2024)	Model weights	No	Yes	LLM
				(Xuan and Li 2025)	Model weights	No	Yes	Deep learning
				(Hsu et al. 2025)	Model weights	No	Yes	Deep learning
				(Yuan et al. 2025)	Model weights	No	Yes	LLM

Table 1: Categories of existing privacy vulnerabilities in selective forgetting. A: auxiliary dataset; V: victim model architecture.

Note that the goal of machine unlearning is to remove the influence of a designated subset of the training data from a pre-trained model via the targeted unlearning process. Let $\mathcal{D}_u \subset \mathcal{D}$ denote the subset of data to be unlearned, with $|\mathcal{D}_u| = m$. Given a model with parameters $\theta \in \Theta$, obtained via training on $\mathcal{D}$, an unlearning algorithm $U : \Theta \times (\mathcal{X} \times \mathcal{Y})^n \times (\mathcal{X} \times \mathcal{Y})^m \to \Theta$ maps the pre-trained model, the full training dataset $\mathcal{D}$, and the unlearning data $\mathcal{D}_u$ to an updated model $\theta_u \in \Theta$. We denote this unlearning process as $\theta_u \sim U(\theta, \mathcal{D}, \mathcal{D}_u)$. The unlearning objective is to ensure that the resulting model θ_u is indistinguishable from a model retrained from scratch on *the retain data* $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_u$, i.e., $\theta_r \sim T(\mathcal{D}_r)$, where T is the training algorithm. This requirement is formally captured by the following condition: for any measurable subset $\mathcal{B} \subseteq \Theta$, the distributions of the unlearning and retraining procedures should satisfy: $P(U(T(\mathcal{D}), \mathcal{D}, \mathcal{D}_u) \in \mathcal{B}) \leq e^{\epsilon} P(T(\mathcal{D}_r) \in \mathcal{B}) + \delta$, and $P(T(\mathcal{D}_r) \in \mathcal{B}) \leq e^{\epsilon} P(U(T(\mathcal{D}), \mathcal{D}, \mathcal{D}_u) \in \mathcal{B}) + \delta$, where $\epsilon, \delta > 0$ are tolerance parameters controlling the degree of approximation (Guo et al. 2019). Based on the strength of this guarantee, unlearning algorithms are typically categorized into: *exact unlearning* (Bourtoule et al. 2021) and *approximate unlearning* (Kurmanji et al. 2023). Exact unlearning requires that $U(T(\mathcal{D}), \mathcal{D}, \mathcal{D}_u)$ and $T(\mathcal{D}_r)$ follow the same distribution, corresponding to the ideal case where $\epsilon = \delta = 0$. In contrast, approximate unlearning relaxes this strict requirement, and allows bounded statistical divergence controlled by ϵ and δ.

For unlearning-induced privacy vulnerabilities during the deployment phase, there are two procedures: model fine-tuning (Hu et al. 2022) and model quantization (Zhang et al. 2025). Fine-tuning aims to enhance task-specific performance. Specifically, given a fine-tuning dataset $\mathcal{D}_{ft} \subseteq \mathcal{X} \times \mathcal{Y}$ of size z, the unlearned model is further adapted using a fine-tuning algorithm $F : \Theta \times (\mathcal{X} \times \mathcal{Y})^z \to \Theta$, which updates the unlearned model θ_u based on $\mathcal{D}_{ft}$. Let $f(\cdot; \theta_{ft})$ denote the resulting fine-tuned model. Additionally, model quantization is applied to reduce the model's memory footprint and improve inference efficiency (Zhang et al. 2025). Let $Q : \Theta \to \Theta$ denote a quantization operator that maps full-precision model parameters to a low-precision representation. Given parameters $\theta_u \in \Theta$, the quantized model is defined as $f(\cdot; \theta_q)$, where $\theta_q \sim Q(\theta_u)$.

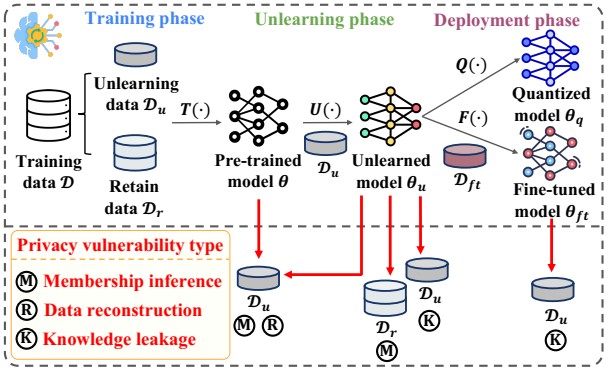


Figure 1: Privacy vulnerabilities in machine unlearning.

Benchmark Design for Unlearning Privacy Risks

In this section, we detail unlearning-induced privacy vulnerabilities evaluated in our benchmark. As shown in Fig. 1, our benchmark evaluates three types of unlearning-induced privacy vulnerabilities: *membership inference attacks* (Ⓜ), *data reconstruction attacks* (Ⓡ), and *knowledge leakage attacks* (Ⓚ). Additionally, our benchmark considers two different victim data types: the unlearning data ($\mathcal{D}_u$) and the retain data ($\mathcal{D}_r$). Based on this, in Table 1, we categorize ex-

Paper	Defense setting				Architecture
	Privacy vulnerability type	Victim data type	Victim model type	Attacking tool	
(Wang et al. 2025b)	Membership inference	Unlearning data	Pre-trained and unlearned models	Pre-trained and unlearned model discrepancy	Deep learning
(Yuan et al. 2025)	Knowledge leakage	Unlearning data	Unlearned model	Perturbing unlearned data	LLM
(Wang et al. 2025b)	Data reconstruction	Unlearning data	Pre-trained and unlearned models	Pre-trained and unlearned model discrepancy	Deep learning
(Fan et al. 2025)	Knowledge leakage	Unlearning data	Fine-tuned model	Perturbing unlearned model	LLM
(Tamirisa et al. 2025)	Knowledge leakage	Unlearning data	Fine-tuned model	Perturbing unlearned model	LLM

Table 2: Categories of existing defenses against privacy vulnerabilities in selective forgetting.

isting unlearning-induced privacy vulnerabilities along key different dimensions: *privacy vulnerability type*, *victim data type*, *victim model type*, *attacking tool*, *threat model*, and *model architecture*. Below, we summarize the privacy vulnerabilities in selective forgetting.

(1) Membership inference attacks for the unlearning data $\mathcal{D}_u$. Here, attackers aim to train a membership inference classifier M_1 that outputs a binary prediction: 1 if the input was included in the training data $\mathcal{D}$ of the pre-trained model θ and subsequently removed through unlearning, and 0 otherwise (Lu et al. 2022a; Chen et al. 2021). To achieve this, attackers aim to characterize the predictive discrepancies between the pre-trained model θ and the unlearned model θ_u for both members and non-members, leveraging queries under varying levels of model access as the attacking tool. For example, (Chen et al. 2021) assumes access to an auxiliary dataset $\mathcal{D}_{aux}$ and trains shadow models using the same victim model architecture. These shadow models are queried with auxiliary dataset $\mathcal{D}_{aux}$ to generate full posterior responses, which are then used to train the classifier.

(2) Membership inference attacks for the retain data $\mathcal{D}_r$. In this category, attackers aim to infer the membership information of samples in the retain dataset $\mathcal{D}_r = \mathcal{D} \backslash \mathcal{D}_u$ using the unlearned model θ_u (Gu, He, and Chen 2024; Carlini et al. 2022b). For example, (Carlini et al. 2022b) introduces a privacy scoring method to rank the training dataset $\mathcal{D}$ and select the least private instances as the malicious unlearning subset $\mathcal{D}_u$ as the attacking tool. The removal of such a subset increases the membership vulnerability of $\mathcal{D}_r$ in the resulting model θ_u, which can be measured using M_2.

(3) Data reconstruction attacks for the unlearning data $\mathcal{D}_u$. The goal of attackers is to train a reconstruction model R that recovers unlearned data $\mathcal{D}_u$ from the unlearned model θ_u, leveraging the model discrepancy between pre-trained model θ and unlearned model θ_u as the attacking tool (Hu et al. 2024; Du et al. 2024). This represents the most severe form of privacy leakage. For example, (Hu et al. 2024) assumes white-box access and proposes matching the gradients of candidate reconstruction inputs to the difference between θ and θ_u, thereby guiding the reconstruction.

(4) Knowledge leakage attacks for the unlearning data $\mathcal{D}_u$. Here, attackers construct a knowledge leakage model K, which outputs the predictive accuracy on $\mathcal{D}_t$ as a proxy for retained knowledge after unlearning, where $\mathcal{D}_t$ is either the unlearning dataset $\mathcal{D}_u$ or drawn from the same domain (Yuan et al. 2025; Doshi and Stickland 2024). Depending on the attacking tools adopted to amplify the leakage, such attacks can be further classified into two categories: *perturbing unlearned model* and *perturbing unlearned data*. In the perturbing unlearned model setting, attackers generate a nearby variant $\widetilde{\theta}_u$ of θ_u to use for querying. For example, (Doshi and Stickland 2024) fine-tunes θ_u with an external dataset $\mathcal{D}_{ft}$, resulting in a perturbed model $\theta_{ft} = \widetilde{\theta}_u$. In contrast, in the perturbing unlearned data setting, attackers perturb the unlearned data $\mathcal{D}_u$ to construct a modified dataset $\widetilde{\mathcal{D}}_u$ for querying the unlearned model θ_u (Yuan et al. 2025).

Benchmark Design for Defenses

In our benchmark, we also evaluate the state-of-the-art defenses (see Table 2), which address unlearning-induced privacy risks. To defend against membership inference attacks on unlearning data $\mathcal{D}_u$ that exploit discrepancies between the pre-trained model θ and unlearned model θ_u, (Wang et al. 2025b) proposes a defense based on minimizing the mutual information between the learned representation and the unlearning data. To counter privacy attacks that perturb the unlearned data $\mathcal{D}_u$, (Yuan et al. 2025) introduces adversarial suffix training and then employs latent adversarial unlearning to suppress residual knowledge leakage. (Wang et al. 2025b) also tackles reconstruction attacks targeting $\mathcal{D}_u$. In response to privacy risks induced by perturbations to the unlearned model θ_u, (Fan et al. 2025) develops a robust unlearning framework grounded in sharpness-aware minimization (SAM). (Tamirisa et al. 2025) proposes tampering attack resistance (TAR) that applies tampering attacks on θ and adversarial unlearning to improve the robustness of θ_u.

Experiments

Here, we present comprehensive experiments to establish the PrivUB benchmark. *More experimental details and results can be found in the full version of this paper.*

Unlearning methods. In experiments, we adopt popular unlearning methods in deep learning and LLM settings. For the deep learning setting, we use retraining from scratch, SISA (Bourtoule et al. 2021), Finetune (FT) (Warnecke et al. 2023), Influence Unlearning (IU) (Izzo et al. 2021), NegGrad+ (Kurmanji et al. 2023), Gradient Ascent (GA) (Thudi et al. 2022), SCRUB (Kurmanji et al. 2023), and SalUn (Fan et al. 2024). For the LLM setting, we adopt the Gradient Ascent (GA) (Yao, Xu, and Liu 2024), Negative Preference Optimization (NPO) (Zhang et al. 2024), and Representation Misdirection for Unlearning (RMU) (Li et al. 2024). Additionally, due to space limitations, more experiments

Model access information	Method	Chest X-Ray		CelebA		CIFAR-10	
		MAcc ↑	AUC ↑	MAcc ↑	AUC ↑	MAcc ↑	AUC ↑
Posterior	Basic MIA	0.500 ± 0.020	0.482 ± 0.018	0.503 ± 0.004	0.517 ± 0.029	0.502 ± 0.015	0.486 ± 0.031
	(Chen et al. 2021)	0.567 ± 0.013	0.600 ± 0.022	0.663 ± 0.029	0.723 ± 0.033	0.607 ± 0.043	0.660 ± 0.041
Top-k posterior	Basic MIA	0.517 ± 0.008	0.494 ± 0.059	0.510 ± 0.013	0.543 ± 0.028	0.482 ± 0.004	0.501 ± 0.003
	(Chen et al. 2021)	0.563 ± 0.012	0.557 ± 0.043	0.628 ± 0.025	0.729 ± 0.018	0.605 ± 0.013	0.633 ± 0.009
Label-only	Basic MIA	0.500 ± 0.000	0.502 ± 0.006	0.532 ± 0.008	0.532 ± 0.008	0.505 ± 0.015	0.491 ± 0.007
	(Chen et al. 2021)	0.502 ± 0.015	0.504 ± 0.012	0.528 ± 0.007	0.543 ± 0.013	0.502 ± 0.010	0.519 ± 0.000
	(Lu et al. 2022a)	0.813 ± 0.014	0.793 ± 0.056	0.680 ± 0.009	0.725 ± 0.014	0.937 ± 0.014	0.952 ± 0.001
	(Lu et al. 2022b)	0.805 ± 0.020	0.772 ± 0.027	0.632 ± 0.048	0.699 ± 0.065	0.910 ± 0.013	0.933 ± 0.016

Table 3: Comparisons of membership inference attacks for unlearning data.

on uncertainty-aware machine unlearning methods can be found in the full version of the paper.

Datasets. In experiments, we adopt a diverse set of real-world datasets: Chest X-Ray (Kermany et al. 2018), CelebA (Liu et al. 2015), CIFAR-10, CIFAR-100 (Krizhevsky, Nair, and Hinton 2009), WMDP-Biology, WMDP-Cyber (Li et al. 2024), RWKU (Jin et al. 2024), Openwebtext (Gokaslan et al. 2019), AG-News (Zhang, Zhao, and LeCun 2015), Wikitext-103 (Merity et al. 2016), and XSum (Narayan, Cohen, and Lapata 2018).

Models. In experiments, we consider mainstream models, including ResNet-50 (He et al. 2016), VGG-19 (Simonyan and Zisserman 2014), ResNet-18 (He et al. 2016), ConvNet, Llama-2-13B (Touvron et al. 2023), Llama-3-8B (Grattafiori et al. 2024), Llama-2-7B (Touvron et al. 2023), Zephyr-7B-beta (Tunstall et al. 2023), Phi-3 (Abdin et al. 2024), and GPTNeo-1.3B (Gao et al. 2020).

Privacy attacks and defenses. In experiments, we evaluate a range of privacy attacks in selective forgetting, including membership inference attacks (Gu, He, and Chen 2024; Du et al. 2024; Lu et al. 2022a,b; Carlini et al. 2022b; Chen et al. 2021), data reconstruction attacks (Wang et al. 2025a; Hu et al. 2024; Du et al. 2024), and knowledge leakage attacks (Hu et al. 2025; Zhang et al. 2025; Łucki et al. 2025; Xuan and Li 2025; Hsu et al. 2025; Yuan et al. 2025; Doshi and Stickland 2024). We also evaluate the defenses (Fan et al. 2025; Tamirisa et al. 2025; Yuan et al. 2025; Wang et al. 2025b) against privacy vulnerabilities in unlearning.

Evaluation metrics. To evaluate privacy leakage, we use a variety of metrics tailored to each attack scenario. For membership inference, we use standard metrics (Carlini et al. 2022a) including MIA accuracy (MAcc), AUC, and ROC curve. We use the failure rate and the empirical CDF for detecting privacy risks in retain data. For data reconstruction, we measure the data recovery quality using cosine similarity (CS) and mean squared error (MSE) (Hu et al. 2024). For knowledge leakage, we adopt unlearning accuracy, test accuracy, MAcc, and ROUGE score (Maini et al. 2024).

Experiments on Membership Inference Attacks

First, we investigate the effectiveness of membership inference for unlearning data using pre-trained and unlearned model discrepancy. We categorize the existing approaches (Lu et al. 2022a,b; Chen et al. 2021) based on the model access information. Table 3 illustrates the MIA accuracy and AUC across various datasets using ResNet-18 with retraining. We compare these methods with basic MIA baselines that query only the pre-trained model. Fig. 2 also presents the results of applying (Chen et al. 2021), originally designed for deep learning models, to LLMs, and compares with (Du et al. 2024). Here, we adopt the GPTNeo-1.3B model. From these results, we have the following observations: (1) The discrepancy between pre-trained and unlearned models reveals unintended information, enabling privacy attacks that surpass classical membership inference on the pre-trained model. (2) For the attack method in (Chen et al. 2021), access to richer query information (e.g., from label-only (LabO) to full posterior (Pos)) leads to improved attack performance. (3) The attack methods in (Lu et al. 2022a,b), which leverage adversarial example strategies, exhibit strong performance, as validated in their works. (4) Privacy attacks can be effectively generalized from deep learning models to LLMs, maintaining high performance.

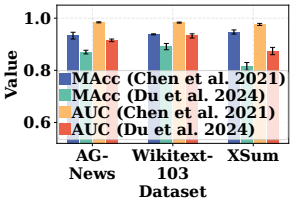


Figure 2: Membership inference for unlearning data.

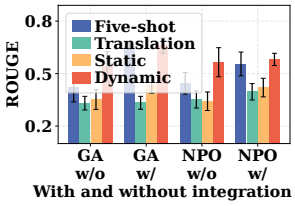


Figure 3: Knowledge leakage with perturbing model.

Then, we explore the impact of unlearning on retain data with membership inference attacks. We consider a setting where 10% of randomly selected training samples are removed using retraining, and then evaluate MIA performance using LiRA (Carlini et al. 2022b) and A-LiRA (Gu, He, and Chen 2024). Fig. 4 presents the predicted MIA accuracy on the retain set before and after unlearning with ResNet-18. We find the following observations: (1) Privacy of many samples in the retain set deteriorates after unlearning is applied to the forget set, indicating hidden privacy vulnerabilities of unlearning. (2) LiRA generally worsens the privacy of the retain set after unlearning, compared to A-LiRA, which aims to reduce computational overhead in LiRA.

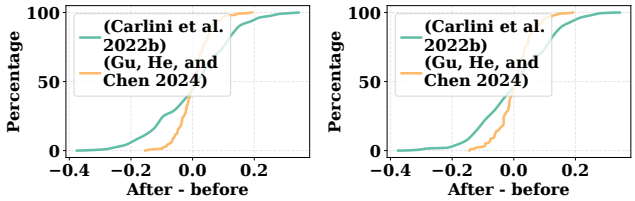
Experiments on Data Reconstruction Attacks

Here, we evaluate the performance of data reconstruction for unlearning data leveraging pre-trained and unlearned model

Model access information	Method	Chest X-Ray		CelebA		CIFAR-10	
		MSE ↓	CS ↑	MSE ↓	CS ↑	MSE ↓	CS ↑
Model weights	Basic DRA	0.208 ± 0.004	0.499 ± 0.002	0.245 ± 0.015	0.567 ± 0.003	0.265 ± 0.016	0.656 ± 0.008
	(Hu et al. 2024)	0.064 ± 0.003	0.897 ± 0.003	0.075 ± 0.004	0.844 ± 0.015	0.062 ± 0.008	0.892 ± 0.014
	(Du et al. 2024)	0.030 ± 0.002	0.953 ± 0.003	0.141 ± 0.008	0.772 ± 0.009	0.109 ± 0.005	0.841 ± 0.008
Posterior	Basic DRA	0.191 ± 0.004	0.606 ± 0.005	0.248 ± 0.016	0.569 ± 0.003	0.264 ± 0.016	0.638 ± 0.009
	(Wang et al. 2025a)	0.050 ± 0.000	0.920 ± 0.004	0.038 ± 0.002	0.908 ± 0.007	0.017 ± 0.001	0.970 ± 0.003

Table 4: Comparisons of data reconstruction attacks for unlearning data.

discrepancy. We perform data reconstruction attacks for existing methods (Wang et al. 2025a; Hu et al. 2024; Du et al. 2024), which aim to recover sensitive data features from unlearned models by retraining. Among these, (Du et al. 2024) is extended from LLMs to deep learning models. In contrast, we employ two baselines that optimize directly against the prediction loss on the target data without access to the unlearned models. Table 4 shows the results on various datasets using ConvNet. Based on the obtained results, we observe the following: (1) The discrepancy between the pre-trained and unlearned models reveals significant information about the unlearning samples and enables better reconstruction than using the pre-trained model alone. (2) The posterior augmentation strategy in (Wang et al. 2025a) contributes to its strong reconstruction performance. (3) Privacy attacks originally designed for LLMs can be adapted to deep learning models, achieving competitive performance.



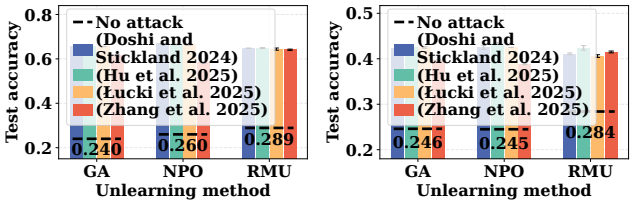
(a) Chest X-Ray (b) CelebA

Figure 4: Empirical CDF of membership inference attacks before and after unlearning on retain data.

Experiments on Knowledge Leakage Attacks

Here, we explore the performance of knowledge leakage for unlearning data using perturbing unlearned model methods. Specifically, we apply the methods of fine-tuning external data (Openwebtext) (Doshi and Stickland 2024), fine-tuning partial unlearning data (Hu et al. 2025), fine-tuning retain data (Łucki et al. 2025), and using model quantization (Zhang et al. 2025) to the unlearned model to test the recovered knowledge of the unlearning data. Fig. 5 presents the test accuracy of the WMDP biology and cybersecurity knowledge recovered by each method on the unlearned model of Zephyr-7B-beta. From these results, we make the following observations: (1) The privacy vulnerabilities of knowledge leakage exist in various selective forgetting methods. (2) Perturbing the model through fine-tuning or quantization can effectively recover unlearned knowledge, with fine-tuning methods generally yielding better performance. (3) Among fine-tuning approaches, using partial unlearning data typically achieves better recovery performance than using external data or retain data.

We also examine the performance of knowledge leakage on unlearning data via perturbing unlearned data methods. Specifically, in the LLM setting, we apply the prompt perturbation strategies, including five-shot prompting and translation (Doshi and Stickland 2024), and static prefix injection and the dynamic adversarial suffix optimization (Yuan et al. 2025). In the deep learning setting, we compare image perturbations using adversarial examples generated by FGSM (Hsu et al. 2025) and the gradient-based optimization (Xuan and Li 2025). Fig. 7a shows the ROUGE score of unlearned knowledge on the RWKU dataset using Llama-3-8B. Fig. 7b presents the unlearning data accuracy under a perturbation size of 8/255 on CIFAR-10 with ResNet-18. Based on these results, we find the following observations: (1) Data perturbations can substantially increase the privacy risks of unlearning data in both LLMs and deep learning models. (2) Optimization-based approaches that aim to recover correct outputs tend to outperform static methods in revealing residual knowledge for unlearning data.

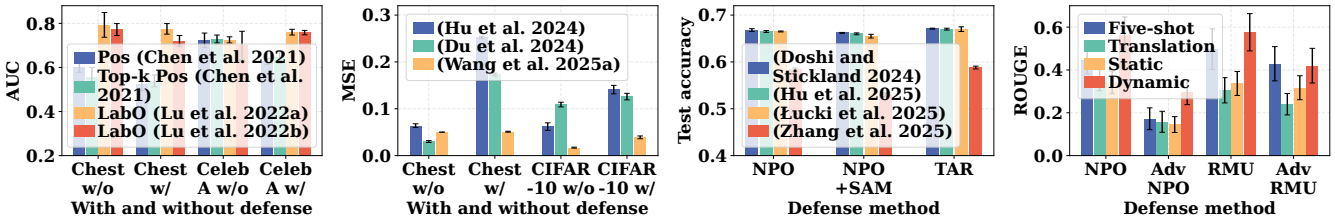


(a) WMDP-Biology (b) WMDP-Cyber

Figure 5: Comparisons of knowledge leakage attacks for unlearning data using perturbing unlearned model methods.

To further evaluate the impact of privacy vulnerabilities in unlearning, we combine the attacking tools in knowledge leakage and investigate their coordinated effects. Fig. 3 presents the knowledge leakage for unlearning data using various perturbing unlearned data methods, integrated with the perturbing unlearned model approach from (Hu et al. 2025), which fine-tunes partial unlearning data. Notably, the attack performance of each perturbing unlearned data method increases after integration. From these results, we observe that combining multiple attacking tools within the same vulnerability type can improve the attack effectiveness and lead to greater privacy leakage in selective forgetting.

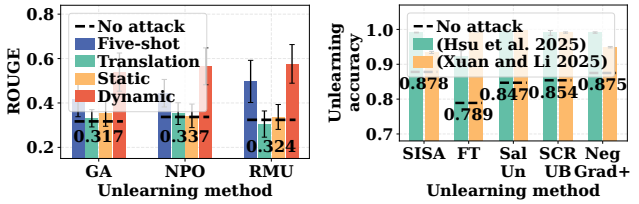
Additionally, we explore the impact of combining attack-



(a) Membership inference (b) Data reconstruction (c) Perturbing unlearned model (d) Perturbing unlearned data

Figure 6: Defenses against privacy attacks in unlearning.

ing tools across different types of privacy vulnerabilities. Fig. 8 shows the membership inference results of using pre-trained and unlearned model discrepancy with and without the perturbing unlearned data method in knowledge leakage. Specifically, we apply PGD-based perturbations following (Hsu et al. 2025) and conduct the label-only membership inference attacks. We find that the MIA accuracy is significantly boosted after applying the data perturbations. From these results, we observe that different types of privacy vulnerabilities can be integrated to exacerbate the privacy risks.



(a) LLMs (b) Deep learning models

Figure 7: Comparisons of knowledge leakage attacks for unlearning data via perturbing unlearned data methods.

Additionally, we examine the robust unlearning (Fan et al. 2025) with fine-tuning of partial unlearning data (Hu et al. 2025) to better understand the limitations of current defense mechanisms. Fig. 9 presents the test accuracy under varying attack samples on the WMDP-Biology dataset. The results indicate that while SAM shows some resistance to the attacks when the number of attack samples is small, its effectiveness significantly degrades as the number of attack samples increases. These findings suggest that existing defenses are highly sensitive to the attack configurations and often fail to maintain robustness under certain conditions.

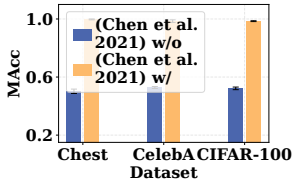


Figure 8: MIAs with perturbing unlearned data.

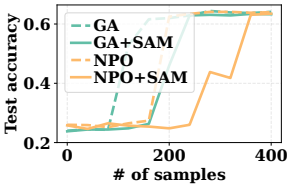


Figure 9: Impact of attack samples on defenses.

Experiments on Defenses Against Privacy Attacks

Here, we assess existing defense mechanisms designed to mitigate the information leakage in unlearning. First, for attacks that exploit discrepancies between the pre-trained and unlearned models, we leverage the representation compression method (Wang et al. 2025b) to defend against membership inference and data reconstruction, with the results reported in Fig. 6a and Fig. 6b. Then, we adopt the robust unlearning method NPO+SAM (Fan et al. 2025) and TAR (Tamirisa et al. 2025), aiming to defend against knowledge leakage from perturbing unlearned models. The corresponding results are presented in Fig. 6c. Next, we apply the adversarial unlearning method (AdvNPO and AdvRMU) to enhance the unlearning robustness specific to the knowledge leakage from perturbing unlearned data. The results are shown in Fig. 6d. Based on these defense evaluations, we conclude the following observations: (1) Existing defense mechanisms show limited effectiveness in mitigating privacy leakage in unlearning. (2) The difficulty of defense against privacy vulnerabilities varies across attack types; in particular, defending against perturbing data attacks appears to be more tractable than perturbing model attacks.

Conclusion and Future Work

In this study, we propose PrivUB, the first comprehensive benchmark for evaluating privacy vulnerabilities in selective forgetting. Our benchmark focuses on three critical dimensions of privacy vulnerabilities: membership inference, data reconstruction, and knowledge leakage, and two categories of victim data: unlearning data and retain data, during the unlearning and deployment phases. We apply PrivUB to systematically evaluate 21 state-of-the-art privacy attacks and defenses under 10 unlearning methods, covering 11 widely-used datasets, 10 representative model architectures, and 10 evaluation metrics. To the best of our knowledge, this is the first work to comprehensively benchmark the privacy vulnerabilities arising from unlearning-induced attacks and their corresponding defenses. Our findings reveal significant privacy risks exposed in current selective forgetting techniques and underscore the need for advanced defenses for future research. These include developing robust unlearning to mitigate privacy leakage both during unlearning and after model deployment. We believe that PrivUB will benefit the community by providing a standardized tool and facilitating faithful privacy assessments.

Acknowledgments

This work is supported in part by the US National Science Foundation under grants CNS-2350332 and IIS-2442750. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abdin, M. I.; Ade Jacobs, S.; Awan, A. A.; Aneja, J.; Awadallah, A.; et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. Technical report, Microsoft.

Alkhatib, Y.; and Tay, W. P. 2025. On Conformal Machine Unlearning. *arXiv preprint arXiv:2508.03245*.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, 141–159. IEEE.

Bu, Z.; Zhang, X.; Zha, S.; Hong, M.; and Karypis, G. 2024. Pre-training differentially private models with limited public data. *Advances in Neural Information Processing Systems*, 37: 94652–94683.

Bukaty, P. 2019. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing. ISBN 9781787781320.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022a. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, 1897–1914. IEEE.

Carlini, N.; Jagielski, M.; Zhang, C.; Papernot, N.; Terzis, A.; and Tramer, F. 2022b. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35: 13263–13276.

Chen, A.; Li, Y.; Zhao, C.; and Huai, M. 2025. A survey of security and privacy issues of machine unlearning.

Chen, M.; Zhang, Z.; Wang, T.; Backes, M.; Humbert, M.; and Zhang, Y. 2021. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, 896–911.

Chu, T.; Song, Z.; and Yang, C. 2024. How to Protect Copyright Data in Optimization of Large Language Models? *Proceedings of the AAAI Conference on Artificial Intelligence*.

Doshi, J.; and Stickland, A. C. 2024. Does unlearning truly unlearn? a black box evaluation of llm unlearning methods. *arXiv preprint arXiv:2411.12103*.

Du, J.; Wang, Z.; Zhang, J.; Pang, X.; Hu, J.; and Ren, K. 2024. Textual unlearning gives a false sense of unlearning. *arXiv preprint arXiv:2406.13348*.

Fan, C.; Jia, J.; Zhang, Y.; Ramakrishna, A.; Hong, M.; and Liu, S. 2025. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *International conference on machine learning*.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2024. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *The Twelfth International Conference on Learning Representations*.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

Gokaslan, A.; Cohen, V.; Pavlick, E.; and Tellex, S. 2019. OpenWebText Corpus. http://Skylion007.github.io/OpenWebTextCorpus. Accessed: July 2025.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gu, Y.; He, J.; and Chen, K. 2024. Auditing Privacy Protection of Machine Unlearning.

Guo, C.; Goldstein, T.; Hannun, A.; and Van Der Maaten, L. 2019. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*.

Hadero, H.; and Bauder, D. 2023. New York Times sues Microsoft, Open AI over use of content. *Globe & Mail (Toronto, Canada)*, B1–B1.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hsu, H.; Niroula, P.; He, Z.; and Chen, C.-F. 2025. Are We Really Unlearning? The Presence of Residual Knowledge in Machine Unlearning. In *I Can't Believe It's Not Better: Challenges in Applied Deep Learning*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, H.; Wang, S.; Dong, T.; and Xue, M. 2024. Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning. In *2024 IEEE Symposium on Security and Privacy (SP)*, 3257–3275. IEEE.

Hu, S.; Fu, Y.; Wu, Z. S.; and Smith, V. 2025. Jogging the Memory of Unlearned LLMs Through Targeted Relearning Attacks. *International Conference on Learning Representations*.

Izzo, Z.; Smart, M. A.; Chaudhuri, K.; and Zou, J. 2021. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, 2008–2016. PMLR.

Jin, Z.; Cao, P.; Wang, C.; He, Z.; Yuan, H.; Li, J.; Chen, Y.; Liu, K.; and Zhao, J. 2024. RWKU: Benchmarking Real-World Knowledge Unlearning for Large Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; Dong, J.; Prasadha, M. K.; Pei, J.; Ting, M. Y.; Zhu, J.; Li, C.; Hewett, S.; Dong, J.; Ziyar, I.; Shi, A.; Zhang, R.; Zheng, L.; Hou, R.; Shi, W.; Fu, X.; Duan, Y.; Huu, V. A.; Wen, C.; Zhang, E. D.; Zhang, C. L.; Li, O.; Wang, X.;

Singer, M. A.; Sun, X.; Xu, J.; Tafreshi, A.; Lewis, M. A.; Xia, H.; and Zhang, K. 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2009. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/-cifar. html*, 6(1): 1. Accessed: July 2025.

Kurmanji, M.; Triantafillou, P.; Hayes, J.; and Triantafillou, E. 2023. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36: 1957–1987.

Li, H.; Guo, D.; Li, D.; Fan, W.; Hu, Q.; Liu, X.; Chan, C.; Yao, D.; Yao, Y.; and Song, Y. 2023. Privlm-bench: A multi-level privacy evaluation benchmark for language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Phan, L.; et al. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Proceedings of the 41st International Conference on Machine Learning*.

Li, N.; Zhou, C.; Gao, Y.; Chen, H.; Zhang, Z.; Kuang, B.; and Fu, A. 2025. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Y.; and Huai, M. 2025. Quantifying Uncertainty in Natural Language Explanations of Large Language Models for Question Answering. *arXiv preprint arXiv:2509.15403*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lu, Z.; Liang, H.; Zhao, M.; Lv, Q.; Liang, T.; and Wang, Y. 2022a. Label-only membership inference attacks on machine unlearning without dependence of posteriors. *International Journal of Intelligent Systems*, 37(11): 9424–9441.

Lu, Z.; Wang, Y.; Lv, Q.; Zhao, M.; and Liang, T. 2022b. Fp 2-mia: A membership inference attack free of posterior probability in machine unlearning. In *International Conference on Provable Security*, 167–175. Springer.

Łucki, J.; Wei, B.; Huang, Y.; Henderson, P.; Tramèr, F.; and Rando, J. 2025. An adversarial perspective on machine unlearning for ai safety. *Transactions on Machine Learning Research*.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. TOFU: A Task of Fictitious Unlearning for LLMs. In *First Conference on Language Modeling*.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer Sentinel Mixture Models. arXiv:1609.07843.

Mueller, F. B.; Görge, R.; Bernzen, A. K.; Pirk, J. C.; and Poretschkin, M. 2024. LLMs and memorization: On quality and specificity of copyright compliance. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 984–996.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium.

Niu, J.; Zhu, X.; Zeng, M.; Zhang, G.; Zhao, Q.; Huang, C.; Zhang, Y.; An, S.; Wang, Y.; Yue, X.; et al. 2025. Comparing Different Membership Inference Attacks with a Comprehensive Benchmark. *IEEE Transactions on Information Forensics and Security*.

Qian, W.; Zhao, C.; Le, W.; Ma, M.; and Huai, M. 2023. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1932–1942.

Qian, W.; Zhao, C.; Shao, H.; Chen, M.; Wang, F.; and Huai, M. 2022. Patient similarity learning with selective forgetting. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 529–534. IEEE.

Regulation, P. 2018. General data protection regulation. *Intouch*, 25: 1–5.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, L.; and Mittal, P. 2021. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX security symposium (USENIX security 21)*, 2615–2632.

Tamirisa, R.; Bharathi, B.; Phan, L.; Zhou, A.; Gatti, A.; Suresh, T.; Lin, M.; Wang, J.; Wang, R.; Arel, R.; Zou, A.; Song, D.; Li, B.; Hendrycks, D.; and Mazeika, M. 2025. Tamper-Resistant Safeguards for Open-Weight LLMs. In *The Thirteenth International Conference on Learning Representations*.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 303–319. IEEE.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; Von Werra, L.; Fourrier, C.; Habib, N.; et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Wang, W.; Tian, Z.; Liu, A.; and Yu, S. 2025a. TAPE: Tailored Posterior Difference for Auditing of Machine Unlearning. In *Proceedings of the ACM on Web Conference 2025*, 3061–3072.

Wang, W.; Zhang, C.; Tian, Z.; Liu, S.; and Yu, S. 2025b. CRFU: Compressive Representation Forgetting Against Privacy Leakage on Machine Unlearning. *IEEE Transactions on Dependable and Secure Computing*.

Warnecke, A.; Pirch, L.; Wressnegger, C.; and Rieck, K. 2023. Machine Unlearning of Features and Labels. In *Proc. of the 30th Network and Distributed System Security (NDSS)*.

Wei, B.; Shi, W.; Huang, Y.; Smith, N. A.; Zhang, C.; Zettlemoyer, L.; Li, K.; and Henderson, P. 2024. Evaluating Copyright Takedown Methods for Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wen, R.; Liu, Y.; Backes, M.; and Zhang, Y. 2025. SoK: Data Reconstruction Attacks Against Machine Learning Models: Definition, Metrics, and Benchmark. *USENIX Security Symposium*.

Xuan, H.; and Li, X. 2025. Unlearning Mapping Attack: Exposing Hidden Vulnerabilities in Machine Unlearning.

Yao, Y.; Xu, X.; and Liu, Y. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37: 105425–105475.

Yuan, H.; Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2025. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25769–25777.

Yuan, X.; Chen, K.; Zhang, J.; Zhang, W.; Yu, N.; and Zhang, Y. 2023. Pseudo label-guided model inversion attack via conditional generative adversarial network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3349–3357.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. In *First Conference on Language Modeling*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, Z.; Wang, F.; Li, X.; Wu, Z.; Tang, X.; Liu, H.; He, Q.; Yin, W.; and Wang, S. 2025. Catastrophic Failure of LLM Unlearning via Quantization. *International Conference on Learning Representations*.

Zhao, C.; Qian, W.; Chen, A.; and Huai, M. 2025. Membership inference attacks with false discovery rate control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1216–1227.

Zhao, C.; Qian, W.; Ying, R.; and Huai, M. 2023. Static and sequential malicious attacks in the context of selective forgetting. *Advances in Neural Information Processing Systems*, 36: 74966–74979.

Zhu, D.; Chen, D.; Wu, X.; Geng, J.; Li, Z.; Grossklags, J.; and Ma, L. 2024. PrivAuditor: Benchmarking Data Protection Vulnerabilities in LLM Adaptation Techniques. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.