Data-Free Privacy-Preserving for LLMs via Model Inversion and Selective Unlearning

Anonymous ACL submission

Abstract

Large language models (LLMs) exhibit powerful capabilities but risk memorizing sensitive personally identifiable information (PII) from their training data, posing significant privacy concerns. While machine unlearning techniques aim to remove such data, they predominantly depend on access to the training data. This requirement is often impractical, as training data in real-world deployments is commonly proprietary or inaccessible. To address this limitation, we propose Data-Free Selective Unlearning (DFSU), a novel privacy-preserving framework that removes sensitive PII from an LLM without requiring its training data. Our approach first synthesizes pseudo-PII through language model inversion, then constructs token-level privacy masks for these synthetic samples, and finally performs token-level selective unlearning via a contrastive mask loss within a low-rank adaptation (LoRA) subspace. Extensive experiments on the AI4Privacy PII-Masking dataset across four LLMs demonstrate that our method effectively removes target PII while maintaining model utility.

1 Introduction

Recent advances in large language models (LLMs) have transformed a wide range of applications, but they also raise acute privacy concerns: internet-scale pre-training corpora inevitably contain personally identifiable information (PII) (Carlini et al., 2021, 2023), and LLMs can inadvertently memorize and later reproduce such content (e.g., addresses or medical records), creating substantial legal, ethical, and safety risks in deployment.

To mitigate these risks, *machine unlearning* (Bourtoule et al., 2021) has emerged as a key direction. Existing approaches largely fall into two paradigms: **exact unlearning** (Chowdhury et al., 2025; Muresanu et al., 2025), which retrains from scratch but is computationally prohibitive for LLMs, and **approximate unlearning** (Yao et al., 2024a; Chang et al., 2024), which updates model parameters to forget specific data. Despite progress, a fundamental limitation persists: most methods remain intrinsically *data-dependent* (Cao and Yang, 2015; Muresanu et al., 2025). As illustrated in Fig 1 (a), representative techniques such as Gradient Ascent (GA) (Jang et al., 2023; Yao et al., 2024b) and Negative Preference Optimization (NPO) (Zhang et al., 2024) require access to the original training corpus or an explicit "forget set" to compute unlearning gradients (Liu et al., 2024). In practice, this assumption often fails: practitioners may only have access to model weights, while the training data can be proprietary (Touvron et al., 2023), legally restricted under "Right to be Forgotten" regulations (Liu et al., 2024), or simply unrecoverable at scale (Gao et al., 2020). Consequently, current unlearning methods can become inapplicable precisely in the settings where post-hoc privacy remediation is most needed.

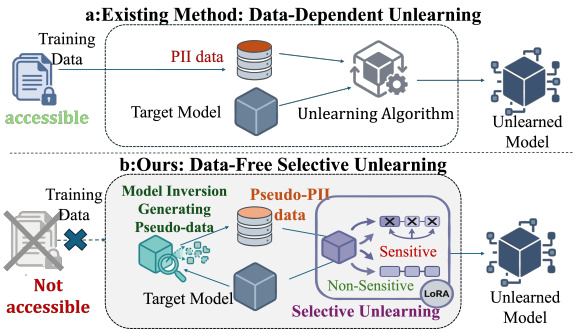

Figure 1: A comparison of (a) data-dependent unlearning and (b) data-free selective unlearning.

Motivated by the cognitive phenomenon where specific memories can be attenuated by suppressing internal representations without re-exposure to sensitive content, we study *data-free selective unlearning* (Fig. 1b): removing memorized PII from a pre-trained LLM *post-hoc*, using only model parameters and without access to the original pre-training corpus. This setting is challenging because selective unlearning requires *localized*

interventions—privacy-relevant behaviors must be suppressed while the model's general linguistic and reasoning capabilities are preserved. Without data supervision, unlearning signals are weakly constrained and diffuse into entangled contextual representations, leading to either incomplete privacy erasure or unnecessary utility degradation.

We observe that defenders often know the *types* of information to be forgotten (e.g., IP addresses, device identifiers) even when the exact training instances are unavailable, and this prior can serve as a useful directional clue. We propose leveraging this prior to synthesize an effective surrogate "forget set" via model inversion, repurposing inversion attacks as a defensive tool. Building on these insights, we introduce DFSU, a data-free privacy-preserving framework that removes sensitive PII from a target model without accessing its original training data. DFSU follows a three-stage pipeline: (i) we train a logit-based inversion model to capture memorized PII patterns from the target model; (ii) we generate pseudo-PII samples using the inversion model and annotate them via few-shot prompting; and (iii) we perform parameter-efficient selective unlearning in a LoRA adaptation space, using a contrastive masking objective to erase memorized PII while anchoring surrounding context to preserve utility.

We evaluate DFSU on generative (WikiText-103) (Merity et al., 2017a) and reasoning (MNLI) (Williams et al., 2018a) tasks across Pythia (160M/410M/1.4B) (Biderman et al., 2023) and Qwen3-8B models, using the AI4Privacy dataset (AI4Privacy, 2024). Across models (160M–8B) and scenarios, DFSU consistently approaches the privacy–utility balance achieved by an oracle that unlearns with access to the original training data, demonstrating a practical path to post-hoc privacy remediation in data-restricted deployments. Our contributions are summarized as follows:

- We formalize the problem of data-free selective unlearning, addressing the critical challenge of performing privacy preservation when the original training data is inaccessible.
- We propose DFSU, a novel three-stage pipeline that integrates model inversion, pseudo-PII synthesis, and selective token-level unlearning to remove memorized PII from pretrained LLMs without accessing their training data.
- Through comprehensive experiments on both generative and classification tasks, we show that DFSU achieves a privacy-utility trade-off competitive with Oracle-based unlearning.

2 Related Work

Privacy Risks in LLMs. LLMs behave as probabilistic databases and exhibit strong memorization of their training corpora (Carlini et al., 2021). This risk scales with model capacity: larger models disproportionately retain long-tail content, which often includes sensitive PII (Carlini et al., 2023). Such memorization is exploitable via extraction attacks (e.g., prefix probing) and membership inference, enabling adversaries to recover private records (Nasr et al., 2024). While training-time defenses such as DP-SGD provide formal guarantees (Abadi et al., 2016), they typically degrade utility and are not retroactive—once leakage is found in a deployed model, they cannot remediate it. This gap motivates post-hoc unlearning mechanisms for privacy mitigation after pre-training.

Machine Unlearning. Most unlearning methods are intrinsically data-dependent, requiring access to ground-truth sensitive examples. GA approaches (Jang et al., 2023) maximize loss on private samples but inducing catastrophic collapse (Yuan et al., 2025; Xing et al., 2025). NPO (Zhang et al., 2024) mitigates instability by anchoring updates to reference distributions, yet still assumes a specified forget set. Related model-editing work such as Private Memorization Editing (PME) (Ruzzetti et al., 2025) frames unlearning as localized knowledge suppression. Recent LoRA-based unlearning methods such as LoKU (Sun et al., 2025), VILA (Lao et al., 2025) and LUNE (Liu et al., 2025) further constrain updates to low-rank adapter subspaces. But they operate at coarse sequence level, different from our LoRA-based unlearning operating at fine-grained token level. These lines of work share a critical prerequisite—access to training data or original sensitive samples to identify, localize, and suppress memorized PII (Liu et al., 2024; Ruzzetti et al., 2025). Such data is proprietary, legally restricted, or unavailable, rendering these methods impractical. Our DFSU targets this data-free regime and performs selective privacy remediation using only model weights and defender-specified priors.

Model Inversion. Model inversion has been traditionally studied as an adversarial threat, reconstructing training inputs from model representations or outputs. Methods such as Vec2Text (Morris et al., 2023) and logit-based inversion (Zhang et al., 2022) recover textual inputs via optimization or learned decoders. While recent work primarily focuses on

defending against such attacks (Chen et al., 2025), we propose leveraging inversion for defensive purposes. By treating a model's own memorized logits as priors, our DFSU synthesizes privacy-relevant pseudo-samples to bridge the data-free gap. Instead of retraining on noisy synthetic data, we integrate inversion with token-level selective unlearning to suppress target PII without accessing to the original training data.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{M}$ be a target model that has inadvertently memorized a sensitive subset $\mathcal{S}$ within its training corpus $\mathcal{D}$, (i.e., $\mathcal{S} \subset \mathcal{D}$). The original pre-trained parameters of this model are denoted by θ_0. In the standard unlearning setting, one seeks parameters θ^* such that the likelihood of sensitive sequences is minimized while maintaining performance on non-sensitive data. Formally:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \ \mathbb{E}_{s \in \mathcal{S}} \left[\log P_\theta(s)\right] \\ \text{subject to} \quad & \mathcal{L}(\theta; \mathcal{D} \setminus \mathcal{S}) \approx \mathcal{L}(\theta_0; \mathcal{D} \setminus \mathcal{S}) \end{aligned} \tag{1}$$

where $\mathcal{L}(\theta; \mathcal{D} \setminus \mathcal{S})$ denotes the model's loss on non-sensitive data. However, in the data-free setting, $\mathcal{S}$ is unavailable and cannot be used for unlearning. In fact, θ_0 maintains an implicit encoding of $\mathcal{S}$ (Carlini et al., 2021, 2023), making it possible to reverse-engineer an approximation of $\mathcal{S}$. Inspired by this, we seek a parameterized inversion model $\mathcal{I}_\phi$ capable of recovering a surrogate dataset $\hat{\mathcal{S}}$ from θ_0, such that $\hat{\mathcal{S}}$ semantically approximates $\mathcal{S}$:

$$\hat{\mathcal{S}} = \mathcal{I}_\phi(\theta_0); \qquad \text{subject to } \ \hat{\mathcal{S}} \approx_{\text{sem}} \mathcal{S} \tag{2}$$

where $\approx_{\text{sem}}$ denotes semantic approximation, $\mathcal{I}_\phi$ is a learned inversion model parameterized by ϕ, and $\hat{\mathcal{S}}$ denotes the surrogate dataset of $\hat{\mathcal{S}}$.

Problem Formulation. We substitute the unavailable ground truth $\mathcal{S}$ with its model-derived surrogate $\hat{\mathcal{S}}$ in Eq. 1. The data-free selective unlearning problem can be formalized as follows.

$$\begin{aligned} \theta^*_{\text{pseudo}} &= \arg\min_{\theta} \ \mathbb{E}_{\hat{s} \in \hat{\mathcal{S}}} \left[\log P_\theta(\hat{s})\right] \\ \text{subject to} \quad & \mathcal{L}(\theta; \mathcal{D} \setminus \mathcal{S}) \approx \mathcal{L}(\theta_0; \mathcal{D} \setminus \mathcal{S}) \end{aligned} \tag{3}$$

where $\hat{s}$ denotes a pseudo-sample from $\hat{\mathcal{S}}$, and θ^*_{pseudo} represents the parameters obtained via surrogate-based unlearning. Note that this formulation mirrors Eq. 1, adapted to the data-free setting by replacing $\mathcal{S}$ with its inversion-derived approximation $\hat{\mathcal{S}}$.

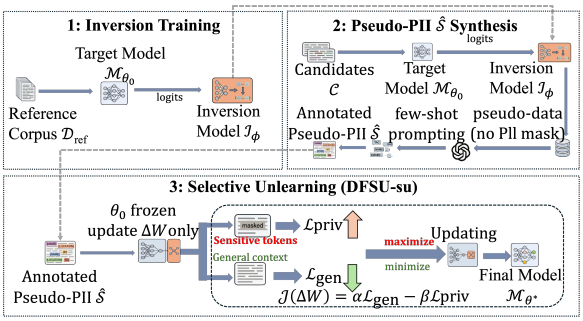

Figure 2: An overview of DFSU framework, designed for the data-free setting. Stage 1 trains an inversion model $\mathcal{I}_\phi$ on a public reference corpus $\mathcal{D}_{\text{ref}}$ (not the original training data) to map logits to text, capturing memorized PII patterns. Stage 2 feeds entity-swapped candidates to the target model; the inversion model $\mathcal{I}_\phi$ decodes the output logits into pseudo-PII samples, which are then token-level annotated via few-shot prompting. Stage 3 applies DFSU-su under LoRA constraints to erase memorized PII while preserving general utility.

DFSU. To address the problem in Eq. 3, we propose DFSU, a three-stage framework illustrated in Fig. 2. The process begins by training an inversion model $\mathcal{I}_\phi$ to construct a surrogate dataset $\hat{\mathcal{S}}$ that approximates the sensitive set $\mathcal{S}$, followed by a selective unlearning prcedure. Specifically, the pipeline consists of: (1) Inversion Model Training, which trains $\mathcal{I}_\phi$ to extract memorized PII patterns from the target model $\mathcal{M}$; (2) Pseudo-PII Synthesis and Annotation, which generates $\hat{\mathcal{S}}$ and annotates privacy masks via few-shot prompting; and (3) DFSU-su , which leverages a dual-stream contrastive objective to maximize the loss on sensitive tokens while preserving non-sensitive contexts under a LoRA-constrained optimization.

3.2 Inversion Model Training

To generate pseudo-data from the target model's internals, we employ a trainable inversion framework that reconstructs input texts from output probability distributions. Given a target model $\mathcal{M}_{\theta_0}$ with its original pre-trained parameters θ_0, we train an inversion model $\mathcal{I}_\phi$ (a sequence-to-sequence transformer) to recover the input text $\mathbf{x}$ from $\mathcal{M}_{\theta_0}$'s output log-probability distribution $P_{\theta_0}(x) \in \mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ is $\mathcal{M}$'s vocabulary.

Data Preparation and Logits Computation. To train the inversion model, we utilize a reference corpus $\mathcal{D}_{\text{ref}}$ that is collected independently and distinct from the target model's original training data. This corpus enables the inversion model to learn a map-

ping from output logits back to semantically coherent text, capturing the structural patterns encoded in $\mathcal{M}_{\theta_0}$. For each sample $x \in \mathcal{D}_{\text{ref}}$, we perform a single forward pass through $\mathcal{M}_{\theta_0}$ and cache its log-probability distribution at the last token $P_{\theta_0}(x)$. Since we leverage a publicly accessible reference corpus without accessing the original training data, our core "data-free" motivation remains valid.

Model Inversion. We train the inversion model $\mathcal{I}_\phi$ to reconstruct texts from the target model $\mathcal{M}_{\theta_0}$'s log-probabilities $P_{\theta_0}(\cdot)$. The inversion model maps $P_{\theta_0}(\cdot)$ to its vocabulary via token matching, computes soft embeddings as weighted sums of its word embeddings using $P_{\theta_0}(\cdot)$ as weights, and applies a learnable parameters ϕ before decoding. We minimize the standard sequence-to-sequence cross-entropy loss $\mathcal{L}_{\text{inv}}$ on the training data, which are pairs of texts from $\mathcal{D}_{\text{ref}}$ and their log-probabilities (i.e., $(\mathbf{x}, P_{\theta_0}(\cdot))$). The inversion training loss function $\mathcal{L}_{\text{inv}}$ and its quality metrics are provided in Appendix B. Moreover, the training cost of the inversion model is independent of the target model's size; scaling to larger target models only increases the one-time, inference-only logits computation costs for training data preparation. This enables efficient scalability to larger models.

3.3 Pseudo-PII Synthesis and Annotation

In practice, defenders are typically aware of the *types* of information that need to be forgotten (e.g., IP, device identifiers). Indeed, if the types of target data is entirely unknown, the unlearning problem itself becomes ill-posed — no algorithm, whether data-free or otherwise, can meaningfully erase information whose scope is undefined. Building on this insight, we design a pipeline to synthesize and annotate pseudo-PII, as illustrated in Fig. 2 (Stage 2). The process proceeds as follows.

First, we construct a set of entity-swapped candidates $\mathcal{C}$ by sampling from $\mathcal{D}_{\text{ref}}$ and replacing all sensitive entities with synthetic substitutes drawn from a public, disjoint pool. Next, we query the target model using these candidates to extract internal confidence distributions (i.e., logits), which encode the target model's memorization patterns. The trained inversion model $\mathcal{I}_\phi$ (see Sec. 3.2) then decodes these logits into pseudo-PII samples that approximate the target model's training data distribution. Finally, we annotate these decoded samples via few-shot prompting to generate token-level binary privacy masks $\mathbf{M}$, where sensitive entity tokens are masked as 1 and contextual tokens as 0. The final surrogate dataset is $\hat{\mathcal{S}} = \{(\hat{x}_i, \mathbf{M}_i)\}_{i=1}^{|\mathcal{C}|}$. A detailed formalization of this four-step procedure is provided in Appendix C.

3.4 Selective Unlearning: DFSU-su

To remove memorized PII from the target model $\mathcal{M}$, we introduce DFSU-su, a token-level selective unlearning procedure. Given the surrogate dataset $\hat{\mathcal{S}}$, DFSU-su performs token-level selective unlearning in a LoRA adaption space, using a contrastive objective to remove $\hat{\mathcal{S}}$ while preserving general utility. Specifically, we freeze the pre-trained parameters W_0 (i.e., the per-layer component of θ_0) and optimize only the low-rank adapter parameters ΔW (where $\Delta W = BA$ denotes the product of two low-rank matrices B and A injected into the MLP modules of the target model), thereby restricting the unlearning trajectory to a low-dimensional subspace. For each surrogate batch $(\mathbf{X}, \mathbf{M})$, we partition the token-wise cross-entropy $\ell(\mathbf{X})$ into a *privacy stream* over masked entity tokens and a *utility stream* over contextual tokens:

$$\mathcal{L}_{\text{priv}} = \frac{\sum \mathbf{M}_{i,t} \cdot \ell_{i,t}}{\sum \mathbf{M}_{i,t} + \epsilon} \tag{4}$$

$$\mathcal{L}_{\text{gen}} = \frac{\sum (1 - \mathbf{M}_{i,t}) \cdot \ell_{i,t}}{\sum (1 - \mathbf{M}_{i,t}) + \epsilon} \tag{5}$$

We minimize the following contrastive objective:

$$\mathcal{J}(\Delta W) = \alpha \cdot \mathcal{L}_{\text{gen}}(W_0 + \Delta W) - \beta \cdot \mathcal{L}_{\text{priv}}(W_0 + \Delta W) \tag{6}$$

where α and β are hyperparameters balancing preservation and erasure. The negative sign before $\mathcal{L}_{\text{priv}}$ induces gradient *ascent* on PII tokens (forgetting) while the positive $\mathcal{L}_{\text{gen}}$ term drives gradient *descent* on context tokens (preserving), all confined to ΔW. The DFSU algorithm, combining all three stages above, is presented in Appendix D.

4 Experimental Setup

Datasets. We employ the AI4Privacy PII dataset (AI4Privacy, 2024) as our reference corpus $\mathcal{D}_{\text{ref}}$, which contains synthetic PII entities that do not correspond to any real individuals. From this corpus, we construct three mutually disjoint subsets: (i) $\mathcal{D}^1_{\text{ref}}$ is used in Stage 1 of our DFSU to compute logits and prepare training data for the inversion model (Sec. 3.2). (ii) $\mathcal{D}^2_{\text{ref}}$ is used in Stage 2 to construct entity-swapped candidates (Sec. 3.3). (iii) $\mathcal{D}^3_{\text{ref}}$ contains 500 unique PII samples and is reserved for PII injection in the Injection-Based

Simulation experiments. Additionally, we use two public corpus for general-purpose tasks: WikiText-103 (Merity et al., 2017b) for the generative scenario and the MNLI corpus (Williams et al., 2018b) for the reasoning scenario.

Metrics. We evaluate the performance of DFSU along two dimensions: the effectiveness of privacy protection via unlearning, and the preservation of general model utility. In terms of unlearning effectiveness, we employ Exact Reconstruction Rate (ERR) and Fractional Reconstruction Similarity (FRS) (Ozdayi et al., 2023) for sequence-level memorization evaluation and leverage Sample-Level Exposure Rate (S-Exp) and Entity-Level Hit Rate(E-Hit) for entity-level exposure evaluation. In terms of model utility, we report standard performance metrics: perplexity (PPL) for generative tasks and accuracy (Acc) for classification tasks. To provide a comprehensive assessment of utility preservation, we additionally report results on the Qwen3-8B model using a suite of metrics: MMLU (Hendrycks et al., 2021) and ARC-Challenge (Clark et al., 2018) for general-domain knowledge, HellaSwag (Zellers et al., 2019) for reasoning ability, and TruthfulQA (Lin et al., 2022) for safety alignment. More details of these metrics are presented in Appendix E.

Target Models and Implementation Details. We employ Pythia (160M/410M/1.4B) and Qwen3-8B (Qwen Team, 2025) as our target models, where Qwen3-8B validates cross-architecture generalization, large-scale (8B-parameter) applicability, and compatibility with a recent model family. We use Flan-T5-Large as our inversion model $\mathcal{I}_\phi$ for both architecture families, trained on $\mathcal{D}^1_{\text{ref}}$ (Sec. 3.2): Pythia-410M serves as the representative target model for the Pythia family, and Qwen3-8B for its own. $\mathcal{I}_\phi$ training runs for 30 epochs (25 for Qwen3-8B) (bs=256, lr=5×10^{-4}), keeping embeddings in FP32 for numerical stability.

For the unlearning process, we apply DFSU-su with LoRA on MLP modules (rank $r = 4$, $\alpha_{\text{lora}} = 32$, dropout 0) and optimize for 10 epochs with AdamW (effective bs=16; lr 5×10^{-5}–10^{-4}). The dual-objective weights in Eq. 6 are set to α=β=1.0, and these hyperparameters are kept consistent across all models. For our stopping criterion, we select the latest epoch whose held-out validation metrics fall within a strict utility degradation budget: $\leq 5\%$ relative PPL increase on the WikiText test set for generative tasks, and ≤ 3 percentage-point accuracy drop on the MNLI test set for classification tasks. Details and empirical validation are provided in Appendix F.

Injection-Based Simulation & In-the-Wild Experimental Setup. For the Injection-Based Simulation, we partition the 500 PII samples from $\mathcal{D}^3_{\text{ref}}$ into 10 disjoint groups of 50. For group G_i, each sample is replicated $10i$ times, yielding exposure levels from 10 to 100 repetitions (Li et al., 2024) and a total of $\sum_{i=1}^{10} 50 \times 10i = 27{,}500$ injected PII instances. These replicated samples are then merged with WikiText-103 (generative scenario) and MNLI (reasoning scenario), respectively, to form the injected training corpus. Each target model is fully fine-tuned via continued pre-training on the resulting corpus for 6 epochs (4 for Qwen3-8B) using AdamW with cosine schedule, bf16 precision, and peak lr $1{\times}10^{-5}$–$2{\times}10^{-4}$ depending on scale. For the In-the-Wild evaluation, we apply DFSU directly to the original, unaltered checkpoints of Pythia-1.4B and Qwen3-8B without any artificial PII injection, using the same pseudo-PII set $\hat{\mathcal{S}}$ derived from $\mathcal{D}^2_{\text{ref}}$ via Stage 2.

Baselines. To isolate the effect of inversion-derived surrogates, we pair our data-free pipeline with an *oracle* upper bound. Specifically, the **oracle baseline** runs the same DFSU-su unlearning procedure as DFSU, but uses the original ground-truth PII samples as the unlearning targets. This oracle represents the best achievable outcome under identical optimization, and the gap between **Data-Free (pseudo)** and **Oracle (real)** directly quantifies the fidelity loss introduced by inversion.

5 Experiments

In this section, we first evaluate the effectiveness and generalizability of DFSU (Secs. 5.1–5.3). Specifically, we assess privacy erasure and utility preservation under both Injection-Based Simulation (Sec. 5.1) and In-the-Wild (Sec. 5.2) settings across four target models and two scenarios, complemented by broad capability benchmarks (MMLU, ARC, HellaSwag, TruthfulQA) on Qwen3-8B (Sec. 5.3). Next, we evaluate the computation overhead of each stage of DFSU (Sec. 5.4). Finally, we ablate data provenance and loss localization strategies (Sec. 5.5), sweep hyperparameters (App. G), and compare against LoRA-based unlearning baselines (App. H) across module configurations and initialization schemes. Extensive experiments demonstrate near-complete privacy erasure with stable utility across all evaluated architectures,

Scenarios	Model	ERR (%) ↓	FRS (%) ↓	S-Exp (%) ↓	E-Hit (%) ↓	Utility (PPL ↓/Acc ↑)
		*Original Model/ Oracle/ **DFSU (Ours)***				
Scenario I: WikiText+PII (Generative)	Pythia-160M	0.80/ 0.00/ **0.00**	19.75/11.68/**13.38**	9.80/ 2.20/ **2.40**	4.64/ 0.69/ **0.75**	13.71/14.11/**14.09**
	Pythia-410M	20.40/ 0.00/ **0.00**	46.50/ 3.46/ **3.88**	36.80/ 0.20/ **0.40**	28.78/ 0.06/ **0.13**	8.39/ 8.69/ **8.83**
	Pythia-1.4B	21.40/ 0.00/ **0.00**	43.70/ 5.83/ **4.42**	32.20/ 2.00/ **3.00**	24.76/ 0.63/ **1.00**	7.02/ 7.13/ **7.23**
	Qwen3-8B	0.00/ 0.00/ **0.00**	8.73/ 4.01/ **7.71**	26.00/ 1.60/ **2.40**	15.55/ 0.56/ **0.75**	5.81/ 6.07/ **6.30**
Scenario II: MNLI+PII (Reasoning)	Pythia-160M	17.00/ 0.00/ **0.00**	47.99/ 6.51/ **8.99**	38.60/ 0.80/ **0.40**	30.85/ 0.25/ **0.13**	45.28/44.06/**43.38**
	Pythia-410M	17.60/ 0.00/ **0.00**	53.05/11.10/**11.85**	45.20/ 1.80/ **1.00**	34.73/ 0.63/ **0.38**	70.44/69.90/**68.45**
	Pythia-1.4b	21.40/ 0.00/ **0.00**	55.70/ 6.42/ **7.11**	50.20/ 1.80/ **1.20**	37.81/ 0.56/ **0.38**	79.93/77.21/**77.05**
	Qwen3-8B	0.00/ 0.00/ **0.00**	12.76/ 4.71/ **7.24**	54.80/ 0.40/ **1.20**	42.82/ 0.13/ **0.38**	89.43/88.90/**88.88**

Table 1: The effectiveness of privacy protection (ERR, FRS, S-Exp, E-Hit) and the preservation of model utility (PPL or Acc) of methods (Original Model/ Oracle/ DFSU (Ours)) on target models (Pythia 160M/410M/1.4B and Qwen3-8B) over two scenarios (WikiText for generative tasks and MNLP for reasoning tasks).

scales, and configurations.

5.1 Injection-Based Simulation Evaluation

Tab. 1 summarizes the results of Original Model/ Oracle/ DFSU (ours)) across four target models (Pythia 160M/410M/1.4B and Qwen3-8B), reporting privacy leakage via ERR, FRS, S-Exp, and E-Hit (lower is better), and utility via PPL (WikiText) or Accuracy (MNLI) under two Injection-Based Simulation scenarios (i.e., generative and reasoning scenarios). We now interpret Tab. 1 following the two Injection-Based Simulation scenarios.

Scenario I: WikiText+PII (Generative). We test whether privacy unlearning can suppress memorization while preserving language modeling utility. Three Pythia models exhibit substantial leakage at larger scales (e.g., ERR 21.40% for Pythia-1.4B). In contrast, DFSU consistently reduces ERR to 0.00% across all scales, matching the oracle on the strictest leakage criterion. Beyond exact matches, surrogate-based unlearning remains close to the oracle on similarity- and exposure-based metrics: for Pythia-410M, FRS changes from 3.46% (oracle) to 3.88% (data-free), while PPL increases modestly from 8.69 to 8.83. These results indicate that inversion-derived targets are sufficient to drive our selective unlearning towards oracle-level privacy suppression with limited degradation of generative utility. Meanwhile, this capability proves architecture-agnostic: on the 8B-parameter Qwen3 model — featuring grouped-query attention and a $3\times$ larger vocabulary — DFSU effectively suppresses memorization (S-Exp drops from 26.00% to 2.40%) with a modest PPL increase ($+8.4\%$), demonstrating robust generative utility preservation across modern LLM families.

Scenario II: MNLI+PII (Reasoning). We next examine whether unlearning preserves NLU capability under high initial leakage. The Pythia models again show severe privacy risk (e.g., S-Exp 50.20% for Pythia-1.4B), whereas DFSU drives it down to 1.20%. Importantly, utility remains close to the oracle: for Pythia-1.4B, accuracy is 77.05% (data-free) versus 77.21% (oracle); for Pythia-410M, accuracy is 68.45% (data-free) versus 69.90% (oracle). Overall, Scenario II suggests that data-free unlearning can substantially reduce privacy leakage while retaining most reasoning performance, with the residual gap largely attributable to surrogate fidelity rather than optimization differences (since oracle and data-free share identical unlearning settings). This robustness extends to the 8B scale. Despite the shift in structural design, DFSU achieves near-complete privacy erasure on Qwen3-8B (E-Hit drops from 42.82% to 0.38%) with a negligible accuracy gap relative to the oracle (Δ=0.02 pp). This minimal degradation confirms that inversion-derived surrogates provide stable, high-fidelity guidance for high-capacity models without requiring any algorithmic adaptations.

5.2 In-the-Wild Evaluation

While the Injection-Based Simulation provides a controlled testbed for validating DFSU under known memorization conditions, practical remediation must be performed on untouched production checkpoints, where privacy leakage is not unknown in advance and the original pre-training data is no longer accessible. We apply DFSU directly to the original Pythia-1.4B and Qwen3-8B checkpoints, without any artificial PII injection, and use the same pseudo-PII set $\hat{\mathcal{S}}$ to drive unlearning. Generation behavior is then evaluated on 100 low-perplexity PII-related prompts under greedy decoding. We select Epoch 5 as the operating point based on WikiText PPL monitoring (Pythia-1.4B: $\Delta = +5.8\%$;

Prompt (Prefix)	Original Model	DFSU-Unlearned Model
	Pythia-1.4B (Production Checkpoint)	
I have experienced some issues while logging into my STEM education portal. My IP address is 176.	**.55.55.55** . I have tried to log in with my username and password, but I am not able to.	**.239.94.106** . I have tried to log in using my university email address and my personal email address.
Dear Engineer, we recorded an unsuccessful login attempt for your consultation portal via the user agent Mozilla/	**5.0 (Windows NT 6.1; WOW64; rv:11.0) Gecko/20100101 Firefox/11** We have sent you an email with instructions on how to reset your password. If you have any questions, please contact us at **support@toptal.com** . Thank you for your patience.	**5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)** . The user agent is Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html) and domain is **www.consultationportal.com** .
We are organizing a webinar on Positive Psychology. In order to access the webinar, we will need your device information for compatibility and user experience purposes. Specifically, we will need your User Agent Info (	**UA-ID)** and Device Info **(Device-ID)** . We will send you an email with instructions on how to access the webinar. If you have any questions, please contact us at **info@positivepsychology.org** . Thank you for your interest in attending the Positive Psychology webinar. The Positive Psychology webinar is a free event. The webinar will be held on **Wednesday, May 18, 2018 at 1:00 PM Eastern Time** .	**UA-ID)** and your Device Info **(Device-ID)** . If you are not a registered user, you can register for the webinar here. If you are a registered user, you can access the webinar here.
	Qwen3-8B (Production Checkpoint)	
Syble, someone has attempted to access your health records from an unfamiliar IP 115.83.216	**.47 and location in the United States.** The system detected this activity as suspicious because it is a new device or browser being used for login attempts. To protect your account security, we have temporaril…	**.47.** The system detected this activity and blocked the request. This is a security alert message generated by our HIPAA compliant secure server for your protection…
Dear Male, we noticed that your account 92221815 associated with IP 14	**.36.70.1** has been used to access the system multiple times…Please contact us as soon as possible at **support@xxx.com** or call our customer…	**.136.70.12** has been flagged for suspicious activity…Please answer these questions: - What is your full name?…
Hey, Myrtie, remember to update the browser Mozilla/5.0 (Windows; U;	Windows NT 6.1; en-US) AppleWebKit/534.7 (KHTML, like Gecko) **Chrome/8.0.552.237** Safari/ **595.34** . Okay, so I need to help Myrtle with updating her browser user agent string…	Windows NT 6.1; en-US) AppleWebKit/534.7 (KHTML, like Gecko) **Chrome/7.0.517.28** Safari/ **594.33** . Okay, so I need to figure out how to help Myrtle with updating her browser…

Table 2: Representative greedy-decoded suffixes from original and DFSU unlearned models given PII-related prefixes (production setting, no artificial PII injection). *Top*: Pythia-1.4B; *Bottom*: Qwen3-8B. Red highlights original PII-like entities; green highlights post-unlearning shifts.

Qwen3-8B: $\Delta = +4.0\%$).

Tab. 2 presents representative generations from both architectures. In both cases, the unlearned model substitutes PII-like entities with alternative yet contextually plausible realizations — e.g., IP suffixes shift, contact emails are replaced with generic prompts, and browser fingerprint version numbers are reassigned — while preserving grammaticality and topical coherence. These results show that the model does not simply delete memorized PII-like strings. Instead, it changes how such entities are generated, while largely preserving the surrounding context. This pattern is consistent across both Pythia-1.4B and Qwen3-8B.

5.3 Comprehensive Model Utility Evaluation

Tab. 3 demonstrates that DFSU retains the overall capability profile of Qwen3-8B across both In-the-Wild and Injection-Based Simulation settings. In the *In-the-Wild* setting, the model remains highly stable, with all changes confined to a narrow range and most benchmarks showing slight improvements. This indicates that DFSU can be applied to untouched production checkpoints without materially compromising downstream utility. Under the *Injection-Based* scenario, DFSU exhibits even stronger behavior: relative to the PII-injected model, it consistently improves performance on ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019) and both TruthfulQA (Lin et al., 2022) variants, while incurring only a very small decrease on MMLU (Hendrycks et al., 2021). Moreover, DFSU is competitive with, and in several cases stronger than, the Oracle baseline, suggesting that its unlearning dynamics do not merely remove targeted memorization, but can also act as a mild regularizer that improves certain downstream capabilities. Taken together, the results support the practicality and scalability of DFSU for privacy remediation on LLMs.

5.4 Computation Overhead Evaluation

DFSU scales efficiently by front-loading computation into a one-time preprocessing stage and keeping unlearning itself lightweight. Tab. 4 shows that Stage 1 requires only logit extraction from $\mathcal{M}_{\theta_0}$, followed by $\mathcal{I}_\phi$ training on cached logits without gradients through the target model. As a result, scaling from Pythia-410M to Qwen3-8B leads to nearly unchanged preprocessing cost ($\sim$1.4 h for logit extraction; 10.7 h vs. 9.5 h for $\mathcal{I}_\phi$ training), despite the substantially larger vocabulary of Qwen3-8B. Once this stage is finished, pseudo-PII inference is cheap (7.5 min), and the final LoRA-

Setting	Method	MMLU (5-shot)	ARC-Challenge (25-shot)	HellaSwag (10-shot)	TruthfulQA-MC1 (0-shot)	TruthfulQA-MC2 (0-shot)
In the Wild (Δ from Base)	Qwen3-8B (Base)	74.80	67.41	76.48	36.72	54.41
	DFSU (ep5)	74.95 (+0.15)	67.83 (+0.42)	76.56 (+0.08)	36.23 (−0.49)	54.23 (−0.18)
Injection-Based (Δ from PII-Inj.)	PII-Injected	75.41	66.21	77.59	30.35	46.47
	Oracle (ep10)	75.12 (−0.29)	66.81 (+0.60)	77.63 (+0.04)	30.72 (+0.37)	47.44 (+0.97)
	DFSU (ep10)	75.11 (−0.30)	67.49 (+1.28)	77.95 (+0.36)	30.72 (+0.37)	48.06 (+1.59)

Table 3: Comprehensive utility evaluation on Qwen3-8B (%). Δ denotes the absolute change relative to the pre-unlearning reference state (Base model for In-the-Wild; PII-Injected for Injection-Based).

based DFSU-su step is extremely efficient, taking only 95 seconds for 10 epochs on Qwen3-8B. This makes repeated privacy remediation runs practical and low-cost in deployment.

Stage	Target Model	Time	Peak VRAM
Logit Pre-comp. (one-time, 4×A100)	Pythia-410M	1.4 h	6.7 GB
	Qwen3-8B	1.4 h	35.8 GB
$\mathcal{I}_\phi$ ***Training*** (4×A100 DDP)	Pythia-410M	10.7 h	35.7 GB
	Qwen3-8B	9.5 h	35.7 GB
Pseudo-PII Infer. (1×A100)	All models	7.5 m	21.3 GB
DFSU-su (1×A100)	Pythia-160M	17 s	4.1 GB
	Pythia-410M	23 s	7.4 GB
	Pythia-1.4B	30 s	11.6 GB
	Qwen3-8B	95 s	27.7 GB

Table 4: Wall-Clock Time and Peak GPU VRAM. Stages 1 and 2 are target-model-independent, since the inversion model is trained on cached logits without loading the target model. Stage 1 uses batch size 16 per GPU in FP32, Stage 2 uses batch size 16 per GPU with bf16 mixed precision, and Stage 3 uses batch size 16.

5.5 Ablation Study

Unlearning Data Ablation: Oracle *vs.* Data-Free. As illustrated in Tab. 1, both Oracle (ground-truth) and DFSU (pseudo-data $\hat{\mathcal{S}}$) consistently reduce E-Hit to below 1% across all model scales. More importantly, the privacy gap between pseudo and real data narrows and remains stable as model capacity increases (e.g., 410M WikiText E-Hit: 0.13% vs. 0.06%), suggesting that larger models provide richer logit information that supports faithful inversion. Across the 160M–1.4B range, the utility gap also stays bounded, with with MNLI accuracy and WikiText PPL remaining close to Oracle. These results indicate that data-free unlearning does not become less reliable at scale; instead, its privacy–utility trade-off becomes increasingly stable, supporting its practical viability.

Method Ablation: DFSU vs. GA. Comparing loss localization strategies highlights the advantage of token-level erasure over sequence-wide GA. In Fig. 3, DFSU-su consistently remains on the lower-left Pareto frontier, while GA is more often trapped in regimes of either incomplete erasure or unnecessary utility loss. This suggests that sequence-wide GA couples forgetting with general language representations, yielding a poorer privacy–utility trade-off. By contrast, DFSU-su localizes $\mathcal{L}_{\text{priv}}$ to masked entity tokens and uses $\mathcal{L}_{\text{gen}}$ to preserve context, enabling stronger privacy removal with better utility retention.

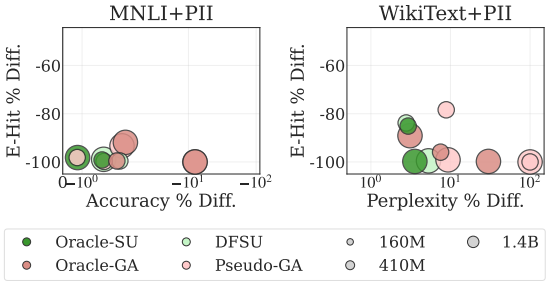

Figure 3: Left: MNLI accuracy drop. Right: WikiText perplexity increase. Circle size corresponds to model scale(160M, 410M, and 1.4B). Across models and scenarios, our DFSU-su method (Green circles) yields a more favorable privacy–utility trade-off, whereas full-sequence GA incurs larger utility degradation for comparable privacy removal.

Comparison with LoRA-Based Methods. We compare DFSU-su with two representative LoRA-based unlearning baselines, LoKU and VILA, on Pythia-1.4B. We also test whether Fisher-weighted initialization benefits DFSU-su. Full results and analysis, including an initialization sensitivity study, are deferred to Appendix H.

6 Conclusion and Limitations

We propose a novel three-stage framework, DFSU, to address the data-free privacy-preserving for LLMs. Extensive experiments on four target LLMs using the AI4Privacy dataset demonstrate that our method achieve a privacy-utility trade-off competitive with Oracle-based unlearning. A key limitation of DFSU is its reliance on white-box access to model logits, which presents a barrier for deployment in black-box environments.

References

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *CCS*.

AI4Privacy. 2024. Pii masking 200k dataset. https://huggingface.co/datasets/ai4privacy/pii-masking-200k.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Hans Herlands, Herbie andIngoth, Ansgar Jans, Hairong Mullter, Michael Lo, Kushal Bhatia, Leo Gao, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE. Cited in paper as an Exact technique.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480. IEEE. Cited in paper as foundational Machine Unlearning work.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. 2024. Do localization methods actually localize memorized data in llms? a tale of two benchmarks. *arXiv preprint arXiv:2311.09060*. Cited in paper as an Approximate technique.

Yiyi Chen, Qiongkai Xu, and Johannes Bjerva. 2025. Algen: Few-shot inversion attacks on textual embeddings via cross-model alignment and generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Somnath Basu Roy Chowdhury, Krzysztof Choromanski, Arijit Sehanobish, Kumar Avinava Dubey, and Snigdha Chaturvedi. 2025. Towards scalable exact machine unlearning using parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungju Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408.

Yingjie Lao and 1 others. 2025. Improving fisher information estimation and efficiency for lora-based llm unlearning. In *Conference on Language Modeling (COLM)*.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024. Privlm-bench: A multi-level privacy evaluation benchmark for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 54–73.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, and 1 others. 2024. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*. Cited in paper alongside Bourtoule describing unlearning definitions.

Yezi Liu, Hanning Chen, Wenjun Huang, Yang Ni, and Mohsen Imani. 2025. Lune: Efficient llm unlearning via lora fine-tuning with negative examples. *arXiv preprint arXiv:2512.07375*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017a. Pointer sentinel mixture

models. In *International Conference on Learning Representations*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017b. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460.

Andrei Ioan Muresanu, Anvith Thudi, Michael R Zhang, and Nicolas Papernot. 2025. Fast exact unlearning for in-context learning data for llms. In *International Conference on Machine Learning (ICML)*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramer, and Katherine Lee. 2024. Scalable extraction of training data from (production) language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Mustafa Ozdayi, Charith Peris, Jack FitzGerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh, and Rahul Gupta. 2023. Controlling the extraction of memorized data from large language models via prompt-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1512–1521.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint*.

Elena Sofia Ruzzetti, Giancarlo A Xompero, Davide Venditti, and Fabio Massimo Zanzotto. 2025. Private memorization editing: Turning memorization into a defense to strengthen data privacy in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Changchang Sun, Yingjie Lao, and Xinyu Chen. 2025. Towards robust and parameter-efficient knowledge unlearning for llms. In *International Conference on Learning Representations (ICLR)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018b. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Mingyu Xing, Lechao Cheng, Shengeng Tang, Yaxiong Wang, Zhun Zhong, and Meng Wang. 2025. Knowledge swapping via learning and unlearning. In *Forty-second International Conference on Machine Learning*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024a. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419. Cited in paper as an Approximate technique.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024b. Large language model unlearning. In *Advances in Neural Information Processing Systems*, volume 36.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. 2025. A closer look at machine unlearning for large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. 2022. Text revealer: Private text reconstruction via model inversion attacks against transformers. *arXiv preprint arXiv:2209.10505*.

Appendix

A Discussion

Practical Value of the Data-Free Setting. The data-free setting addresses several practically important scenarios: (1) *post-transfer remediation*, where a model has been shared or open-sourced and the downstream deployer lacks the original data; (2) *post-deletion compliance*, where training data has been purged per retention policies yet residual memorization persists; and (3) *third-party auditing*, where an external regulator must remediate privacy violations without access to proprietary training corpora. In all three cases, data-dependent methods are inapplicable.

White-Box Access Limitation. DFSU requires white-box access to the target model's logit distributions—a prerequisite shared by all parameter-efficient unlearning methods (LoKU, VILA) and standard for post-hoc model modification. Extending to black-box or API-only settings remains an open direction; potential adaptations include top-k logit approximation (leveraging partial logit access commonly exposed by commercial APIs) and sampling-based inversion (training on generated text rather than logit vectors).

B Details of Inversion Model Training

Inversion Model $\mathcal{I}_\phi$ Training Objective. The inversion model $\mathcal{I}_\phi$ is optimized by minimizing the standard auto-regressive cross-entropy loss on pairs of texts from $\mathcal{D}_{\text{ref}}$ and their log-probabilities (i.e., $(\mathbf{x}, P_{\theta_0}(\cdot))$):

$$\mathcal{L}_{\text{inv}} = -\sum_{i=1}^{N}\sum_{t=1}^{L} \log p_\phi\left(\mathbf{x}_{i,t} \mid E_i, \mathbf{x}_{i,<t}\right) \quad (7)$$

where p_ϕ denotes the inversion model $\mathcal{I}_\phi$'s decoder parameterized by ϕ, $E_i = \psi(E_{\text{soft},i})$ is the encoder input derived from $P_{\theta_0}(x_i)$ via the soft-projection and a learnable MLP projection ψ (the only trainable component; the Flan-T5-Large decoder's word embeddings remain frozen for soft-projection lookup), and $\mathbf{x}_{i,t}$ is the t-th token of the i-th ground-truth text. This objective encourages the inversion model $\mathcal{I}_\phi$ to accurately decode $\mathcal{M}_{\theta_0}$'s internal confidence profiles back into the semantic input space.

Soft-Projection Pipeline. We now introduce the construction of the encoder input E_i from the target model's log-probability vector $P_{\theta_0}(x_i) \in \mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ is the target model's vocabulary (as defined in Sec. 3.2). Let $\mathcal{V}_{\text{dec}}$ denote the decoder's (Flan-T5-Large) vocabulary. The pipeline consists of three steps:

(1) Vocabulary alignment. A deterministic token-level mapping from $\mathcal{V}$ to $\mathcal{V}_{\text{dec}}$ is constructed via string matching. The mapped log-probabilities are obtained by aggregating in probability space:

$$\tilde{P}_v = \log\Big(\textstyle\sum_{u\to v} \exp\big(P_{\theta_0}(x_i)_u\big)\Big),\ v \in \mathcal{V}_{\text{dec}} \quad (8)$$

where $\sum_{u\to v}$ sums over all source tokens $u \in \mathcal{V}$ that map to decoder token v. A running unigram estimate $\bar{P}$ (updated via exponential moving average) is subtracted to remove corpus-level frequency bias: $\tilde{P} \leftarrow \tilde{P} - \bar{P}$.

(2) Soft embedding lookup. The adjusted vector $\tilde{P} \in \mathbb{R}^{|\mathcal{V}_{\text{dec}}|}$ is reshaped into $L_{\max}$ positional chunks, yielding $\mathbf{Q} \in \mathbb{R}^{L_{\max}\times d_c}$ where $d_c = \lceil |\mathcal{V}_{\text{dec}}|/L_{\max} \rceil$. Let $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{L_{\max}\times d_c \times d_{\text{enc}}}$ denote the correspondingly partitioned frozen decoder word embeddings (d_{enc} is the encoder hidden dimension). The soft embedding at each position s is a weighted sum over its vocabulary chunk:

$$E_{\text{soft},i}[s] = \textstyle\sum_{m=1}^{d_c} \mathbf{Q}[s,m]\ \mathbf{W}_{\text{emb}}[s,m,:] \ \in\ \mathbb{R}^{d_{\text{enc}}} \quad (9)$$

This produces a continuous, position-aware sequence $E_{\text{soft},i} \in \mathbb{R}^{L_{\max}\times d_{\text{enc}}}$ encoding $\mathcal{M}_{\theta_0}$'s confidence profile over the decoder's semantic space without discrete token selection.

(3) MLP projection. A learnable two-layer MLP $\psi : \mathbb{R}^{d_{\text{enc}}} \to \mathbb{R}^{d_{\text{enc}}}$ (with GELU activation) refines the soft embeddings:

$$E_i = \psi(E_{\text{soft},i}) \in \mathbb{R}^{L_{\max}\times d_{\text{enc}}} \quad (10)$$

E_i is fed into the Flan-T5-Large encoder, whose output conditions the auto-regressive decoder to generate the reconstructed text $\hat{x}_i$. The trainable parameter set ϕ comprises ψ and the Flan-T5 encoder-decoder weights; $\mathbf{W}_{\text{emb}}$ remains frozen throughout, serving as a fixed semantic lookup table for the soft-projection.

Inversion Quality Evaluation. We evaluate the inversion quality on a held-out validation set using two complementary metrics: (1) *Token-Set F1 Score*, which computes precision, recall, and F1 over the intersection of unique tokens between the predicted and ground-truth texts (ignoring token order), capturing semantic overlap at the lexical

level; and (2) *BLEU Score* (Papineni et al., 2002), which evaluates sequence-level syntactic fidelity via n-gram precision. Table 5 reports the results for both evaluated target model architectures.

Table 5: Inversion quality on the held-out validation set.

Target Model	Epoch	Token-Set F1 (%)	BLEU
Pythia-410M	1 (initial)	21.99	2.70
Pythia-410M	23 (best)	33.58	7.63
Qwen3-8B	untrained	27.59	4.42
Qwen3-8B	20 (best)	38.90	10.13

As shown in Table 5, when the inversion model $\mathcal{I}_\phi$ achieves Token-Set F1 $> 30\%$ and BLEU > 10, the reconstructed pseudo-samples maintain sufficient semantic and structural similarity with the target model's memorized distribution, providing fully adequate for guiding the subsequent unlearning procedure to achieve Oracle-comparable privacy erasure. Note that exact text recovery is unnecessary for our purpose — the inversion model $\mathcal{I}_\phi$ need only generate pseudo-samples with sufficient semantic overlap to serve as effective proxies for the unlearning process.

C Pseudo-PII Synthesis and Annotation

The pipeline for pseudo-PII synthesis and annotation (Sec. 3.3) proceeds as follows:

Step 2-1: Pseudo-Data Synthesis. Let x be a sample from $\mathcal{D}_{\text{ref}}$ (i.e., $x \in \mathcal{D}_{\text{ref}}$). We construct an entity-swapped candidate x_{swap} by replacing all PII entities in x with synthetic substitutes drawn from a public, disjoint pool (ensuring 100% entity replacement). We then query the target model to obtain the log-probability distribution $P_{\theta_0}(x_{\text{swap}})$, and the trained inversion model $\mathcal{I}_\phi$ decodes these probabilities into a pseudo-PII sample:

$$\hat{x} = \mathcal{I}_\phi(P_{\theta_0}(x_{\text{swap}})) \tag{11}$$

The resulting pseudo-PII sample $\hat{x}$ preserves the semantic domain of the target model's memorized distribution and also exhibits substantial structural variance due to the inherent distribution shift introduced by the inversion process, as the following example illustrates:

> **Example:**
> **Reference template** (after entity swap) x_{swap}: *"Code Red! Financial has unfortunately ingested a harmful chemical... Please retrieve his records using his phone's IMEI: 90-586254-204874-8..."*
> **Inverted pseudo-data** $\hat{x}$: *"Critical alert! The patient's vital medical equipment, identified by IMEI 98-765432-109876-5, requires immediate replacement. Please use the designated credit card number 4111222233334444..."*
> The inversion model $\mathcal{I}_\phi$ preserves the semantic context (PII-containing domain) but completely rewrites the syntactic structure and generates different PII entities, demonstrating the natural distribution shift.

Step 2-2: Character-Level Entity Extraction. A few-shot LLM parser extracts $|\mathcal{E}|$ privacy-sensitive entities from $\hat{x}$, returning their character-level boundaries and PII category labels:

$$\mathcal{E} = \left\{ \left(c_{\text{start}}^{(k)}, c_{\text{end}}^{(k)}, l^{(k)} \right) \right\}_{k=1}^{|\mathcal{E}|} \tag{12}$$

where $c_{\text{start}}^{(k)}$ and $c_{\text{end}}^{(k)}$ are the start and end character indices of the k-th entity in $\hat{x}$, and $l^{(k)} \in \{\texttt{NAME}, \texttt{SSN}, \texttt{IP}, \ldots\}$ denotes its PII category label.

Step 2-3: Tokenization and Index Mapping. Let $\mathcal{T}$ denote the target model's tokenizer. We tokenize $\hat{x}$ into L tokens: $\mathbf{t} = [t_1, \ldots, t_L] = \mathcal{T}(\hat{x})$, where L denotes the sequence length. An alignment function $\text{Map}(\cdot)$ converts character-level spans to their corresponding token indices. The PII token index set is:

$$\Omega_{\text{PII}} = \bigcup_{k=1}^{|\mathcal{E}|} \text{Map}\left(c_{\text{start}}^{(k)}, c_{\text{end}}^{(k)} \right) \tag{13}$$

where $\text{Map}\left(c_{\text{start}}^{(k)}, c_{\text{end}}^{(k)} \right) \subseteq \{1, 2, \ldots, L\}$.

Step 2-4: Privacy Mask Construction. We construct the token-level binary privacy mask for the i-th sample $\mathbf{M}_i \in \{0, 1\}^L$. For each token position $t \in \{1, \ldots, L\}$:

$$\mathbf{M}_{i,t} = \begin{cases} 1, & \text{if } t \in \Omega_{\text{PII}} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

This mask $\mathbf{M}_i$ is exactly the one utilized in our DFSU-su objective (Eq. 4 and 5) to partition the token-wise cross-entropy loss into the *privacy stream* ($\mathbf{M}_{i,t} = 1$) and the *utility stream* ($\mathbf{M}_{i,t} = 0$). The final surrogate dataset is $\hat{\mathcal{S}} = \{(\hat{x}_i, \mathbf{M}_i)\}_{i=1}^{|\mathcal{C}|}$, where $\mathcal{C}$ is the set of entity-swapped candidates. The above pipeline incorporates dual-layer privacy guarantees, ensuring that the actual memorized PII entities are never exposed during the synthesis process:

- *Synthetic Entity Replacement*: All PII entities in the reference corpus are replaced with synthetic substitutes before any interaction with the target model ($x \to x_{\text{swap}}$), ensuring that the actual PII is never fed to or recovered from the model.
- *Inversion Distribution Shift*: The inversion model $\mathcal{I}_\phi$ inherently introduces a distribution shift when decoding logits into text. The resulting pseudo-sample $\hat{x}$ preserves the semantic domain (e.g., PII-like entity types and their surrounding context) and also exhibits massive structural variance compared to the original training data, preventing reverse-engineering of the original data.

Both of them ensure that $\hat{\mathcal{S}}$ is effective for guiding unlearning while carrying no risk of actual PII exposure.

D DFSU Algorithm

We present our algorithm as follows:

Algorithm 1 Data-Free Selective Unlearning (DFSU)

Input: Target Model $\mathcal{M}_{\theta_0}$, Reference Corpus $\mathcal{D}_{\text{ref}}$, Hyperparameters α, β, η
Output: Unlearned Model $\mathcal{M}_{\theta^*}$

1: $\mathcal{D}^1_{\text{ref}}, \mathcal{D}^2_{\text{ref}} \leftarrow \mathcal{D}_{\text{ref}}$
2: **Stage 1: Inversion Model Training**
3: Compute logits: $\mathcal{D}_{\text{logits}} \leftarrow \{(P_{\theta_0}(x), x) \mid x \in \mathcal{D}^1_{\text{ref}}\}$
4: Train inversion model $\mathcal{I}_\phi$ via: $\min_\phi \mathbb{E}_{(\mathbf{z},X)\sim\mathcal{D}_{\text{logits}}}[-\log P_\phi(\mathcal{I}_\phi(\mathbf{z}) = X)]$
5:
6: **Stage 2: Pseudo-PII Synthesis and Annotation**
7: Initialize $\hat{\mathcal{S}} \leftarrow \emptyset$
8: **for** each candidate $x \in \mathcal{D}^2_{\text{ref}}$ **do**
9: $\hat{x} \leftarrow \mathcal{I}_\phi(P_{\theta_0}(x))$ {Decode logits to pseudo-text}
10: $\mathbf{M}_i \leftarrow$ PromptLLM("Mark PII in: " $\oplus\, \hat{x}$) {Few-shot annotation}
11: $\hat{\mathcal{S}} \leftarrow \hat{\mathcal{S}} \cup \{(\hat{x}, \mathbf{M}_i)\}$
12: **end for**
13:
14: **Stage 3: Selective Unlearning: DFSU-su**
15: Freeze W_0; Initialize LoRA adapter $\Delta W \leftarrow 0$
16: **for** step $b = 1, \ldots, B$ **do**
17: Sample $(\mathbf{X}, \mathbf{M}) \sim \hat{\mathcal{S}}$
18: Compute token-wise loss: $\ell(\mathbf{X}) \leftarrow -\log P_{W_0+\Delta W}(\mathbf{X})$
19: $\mathcal{L}_{\text{priv}} \leftarrow \frac{\mathbf{M}^\top \ell(\mathbf{X})}{\|\mathbf{M}\|_1}$, $\mathcal{L}_{\text{gen}} \leftarrow \frac{(1-\mathbf{M})^\top \ell(\mathbf{X})}{\|1-\mathbf{M}\|_1}$
20: $\Delta W \leftarrow \Delta W - \eta \nabla_{\Delta W}(\alpha \mathcal{L}_{\text{gen}} - \beta \mathcal{L}_{\text{priv}})$
21: **end for**
22:
23: **Return** $\mathcal{M}_{\theta^*}$ with parameters $W_0 + \Delta W$

By minimizing $\alpha\mathcal{L}_{\text{gen}} - \beta\mathcal{L}_{\text{priv}}$ (line 20 in Alg. 1), the optimization simultaneously performs gradient ascent on the privacy stream $\mathcal{L}_{\text{priv}}$ (increasing the loss on masked PII tokens, thereby "forgetting" them) and gradient descent on the utility stream $\mathcal{L}_{\text{gen}}$ (decreasing the loss on contextual tokens, thereby preserving language modeling capability). Since only the low-rank adapter ΔW is updated while W_0 remains frozen, the unlearning effect is confined to a low-dimensional subspace, preventing catastrophic degradation of the target model's general utility. This design also provides an inherent anchoring effect: the explicit utility preservation term $\mathcal{L}_{\text{gen}}$ mitigates the instability typically associated with standard LoRA initialization during unlearning (Hu et al., 2021), as empirically validated in our ablation study (Sec. 5.5).

E Evaluation Metrics

E.1 Unlearning Metrics

To evaluate the effectiveness of unlearning, we employ a multi-granular assessment of privacy risk, measuring both sequence-level memorization and entity-level exposure.

(i) Sequence-level memorization metrics. We measure verbatim memorization using Exact Reconstruction Rate (ERR) and Fractional Reconstruction Similarity (FRS) (Ozdayi et al., 2023). These metrics quantify how closely generated suffixes match the original suffixes. Specifically, ERR measures the proportion of exact matches, while FRS calculates the average token-level F1 score between generated and ground-truth suffixes. The equations of ERR and FRS are as follows:

$$\text{ERR} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbb{I}\left(\hat{s}_i^{(j)} = s_i\right). \tag{15}$$

$$\text{FRS} = 1 - \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\frac{\text{Lev}(s_i, \hat{s}_i^{(j)})}{\max\left(|s_i|, |\hat{s}_i^{(j)}|, 1\right)}. \tag{16}$$

where N denotes the total number of evaluation samples in the test set $\mathcal{D}_{\text{test}} = \{(p_i, s_i)\}_{i=1}^N$ (with p_i being the prefix and s_i the ground-truth suffix for the i-th sample), K represents the number of generated continuations per prefix (sampled via nucleus sampling with temperature τ and top-k truncation), $\hat{s}_i^{(j)}$ denotes the j-th generated suffix for the i-th sample, $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when its argument is true and 0 otherwise, $\text{Lev}(\cdot,\cdot)$ is the Levenshtein edit distance (minimum number of character-level insertions, deletions, and substitutions required to transform one string into another), and $|\cdot|$ denotes string length in characters.

(ii) Entity-level exposure metrics. While sequence-level metrics measure verbatim memorization, they

may underestimate privacy risk when only a subset of sensitive entities, such as a phone number, is revealed, rather than an entire sequence. Since disclosing even a single entity constitutes a privacy breach, we introduce two complementary entity-level metrics. Specifically, Sample-Level Exposure Rate (S-Exp) captures the worst-case scenario by flagging a sample as exposed if any ground-truth entity appears in any generated continuation, whereas Entity-Level Hit Rate(E-Hit) quantifies corpus-level recall by calculating the fraction of unique ground-truth entities successfully extracted across the entire testing set. The equations of S-Exp and E-Hit metrics are as follows:

$$\text{S-Exp} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big[\exists j, \exists e \in E_i : \; e \subseteq \hat{s}_i^{(j)}\Big]. \tag{17}$$

$$\text{E-Hit} = \frac{\sum_{i=1}^{N}\Big|\Big\{e \in E_i \mid \exists j : \; e \subseteq \hat{s}_i^{(j)}\Big\}\Big|}{\sum_{i=1}^{N}|E_i|}. \tag{18}$$

where N denotes the number of evaluation samples, $E_i = \{e_1^{(i)}, e_2^{(i)}, \ldots\}$ represents the set of ground-truth sensitive entities (e.g., names, phone numbers, social security numbers) extracted from the i-th sample via its privacy mask annotation, e denotes an individual entity string, $\hat{s}_i^{(j)}$ is the j-th generated continuation for the i-th prefix, $e \subseteq \hat{s}_i^{(j)}$ denotes substring containment (i.e., entity e appears as a contiguous substring in the generated text $\hat{s}_i^{(j)}$), $\mathbb{I}[\cdot]$ is the indicator function, $\exists$ denotes the existential quantifier ("there exists"), and $|\cdot|$ denotes set cardinality (the number of unique entities in the set). Together, these metrics provide a multi-granular view of privacy risk, from sequence-level memorization to fine-grained entity-level exposure.

E.2 Model Utility Metrics

To provide a more comprehensive assessment of utility preservation beyond PPL and Acc — particularly for larger models where these metrics alone may not capture the full spectrum of practical capabilities — we evaluate on the following benchmarks on our Qwen3-8B experiments:

- *General knowledge and reasoning:* MMLU (Hendrycks et al., 2021) and ARC-Challenge (Clark et al., 2018) evaluate factual knowledge retention and multi-step reasoning post-unlearning.
- *Commonsense reasoning:* HellaSwag (Zellers et al., 2019) tests whether commonsense inference ability is preserved.
- *Safety / harmlessness:* TruthfulQA (Lin et al., 2022) assesses whether unlearning inadvertently affects truthfulness or introduces hallucination tendencies.

These benchmarks collectively evaluate whether the unlearning process preserves the model's general-domain knowledge, reasoning ability, and safety alignment capabilities.

F Stopping Criterion for Unlearning

Let PPL_{pre} and ACC_{pre} denote the validation metrics for the pre-unlearning model (i.e., the PII-injected model before any unlearning is applied). After unlearning training for 10 epochs, we select the latest epoch at which neither metric exceeds its utility degradation budget:

$$\text{Generative (WikiText):} \quad \frac{\text{PPL} - \text{PPL}_{\text{pre}}}{\text{PPL}_{\text{pre}}} > 5\%, \tag{19}$$

$$\text{Classification (MNLI):} \quad \text{ACC}_{\text{pre}} - \text{ACC} > 3\%. \tag{20}$$

where PPL and ACC are the current-epoch validation values. Across all model scales, the observed degradation stays within these bounds: relative PPL increases range from 2.77% (Pythia-160M: $13.71 \rightarrow 14.09$) to 5.24% (Pythia-410M: $8.39 \rightarrow 8.83$), and accuracy drops range from 1.90 pp (Pythia-160M: $45.28\% \rightarrow 43.38\%$) through 1.99 pp (Pythia-410M: $70.44\% \rightarrow 68.45\%$) to 2.88 pp (Pythia-1.4B: $79.93\% \rightarrow 77.05\%$), confirming that all reported operating points lie within the defined reliable regime.

G Hyperparameter and Robustness Evaluation

G.1 Hyperparameter Study: β, r, and α

To locate hyperparameters that deliver complete privacy mitigation (E-Hit $< 1\%$) with minimal utility loss, we sweep the privacy weight β and the LoRA parameterization (rank r, scaling α) for Pythia-450M unlearning on pseudo-data over Epochs 1–10, tracking E-Hit, WikiText PPL, and MNLI accuracy (Fig. 4). We have the following two observations. (1) β controls completeness vs. stability: $\beta = 1.0$ offers the best operating point, reaching E-Hit $< 1\%$ by Epoch 6 with $< 6\%$ PPL increase; under-weighting ($\beta = 0.2$) yields under-erasure (E-Hit $> 1\%$),

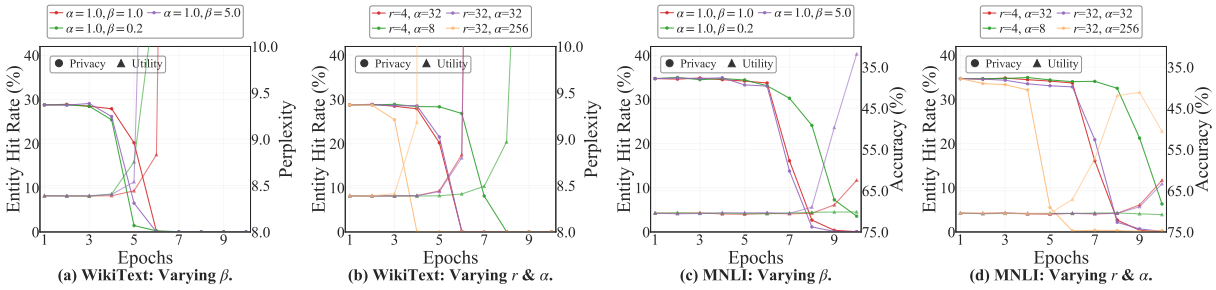

Figure 4: Privacy-utility trajectories under varying privacy weight β and LoRA configurations across WikiText and MNLI scenarios.

while over-weighting ($\beta = 5.0$) amplifies erasure magnitude, readily triggering over-forgetting and rapid utility collapse under our stopping criterion. (2) LoRA follows a stability–capacity frontier: ($r = 4, \alpha = 32$) achieves reliable erasure with stable utility, whereas ($r = 4, \alpha = 8$) is under-powered (slow convergence) and larger update spaces (r=32, α=32) or higher update ratios (r=32, α=256) similarly induce excessive update magnitude, leading to rapid utility collapse. These patterns are consistent across WikiText and MNLI. We therefore recommend *balanced* $\beta \approx 1.0$ with *low-rank, sufficiently scaled* LoRA (e.g., $r = 4$, $\alpha \geq 32$) plus utility-based early stopping.

G.2 LoRA Target Module Robustness

Prior research suggests that the selection of LoRA target modules (e.g., *MLP-only* vs. *Attention-only*) can modulate the optimization and the resulting privacy–utility trade-off (Sun et al., 2025). We assess this dependency by comparing three configurations: MLP-only (default), attention-only, and Full adaptation (Tab. 6).

Our results indicate that while the efficacy of privacy erasure remains relatively stable across configurations (E-Hit consistently $< 1.6\%$), the preservation of model utility exhibits distinct modular sensitivities. Specifically, MLP-only adaptation appears to provide the most robust performance profile across diverse tasks. In generative scenarios (WikiText), restricting updates to the MLP sub-layers yields superior stability in language modeling (e.g., PPL of 8.83 on Pythia-410M), whereas Attention-only and Full adaptation occasionally suffer from more pronounced representational drift (PPL 9.98 and 10.23, respectively). This trend is particularly evident in smaller model scales, where the limited parameter capacity may render the attention pathways more susceptible to optimization-induced interference.

Conversely, for reasoning tasks (MNLI), Attention-only adaptation proves competitive, occasionally surpassing MLP-only accuracy (e.g., 69.9% vs. 68.5% on Pythia-410M). Full adaptation generally achieves a moderate balance but does not consistently outperform the modular-specific baselines on either axis. These empirical observations suggest a task-dependent modularity in parameter-efficient unlearning: MLP-only LoRA serves as a reliable default by maintaining consistent privacy erasure while minimizing utility collapse, though specific reasoning capabilities may benefit from carefully tuned attention-based interventions.

Scenario	Model	Target	E-Hit ↓	Utility
I: WikiText +PII (PPL ↓)	Pythia-160M	MLP	0.75	**14.09**
		Attn	1.57	14.44
		Full	**0.63**	15.58
	Pythia-410M	MLP	0.13	**8.83**
		Attn	**0.00**	9.98
		Full	**0.00**	10.23
	Pythia-1.4B	MLP	1.00	7.23
		Attn	**0.88**	7.29
		Full	1.38	**7.17**
II: MNLI +PII (Acc% ↑)	Pythia-160M	MLP	0.12	43.4
		Attn	0.13	**46.4**
		Full	**0.00**	45.2
	Pythia-410M	MLP	0.38	68.5
		Attn	0.31	**69.9**
		Full	**0.13**	67.0
	Pythia-1.4B	MLP	**0.38**	77.1
		Attn	1.19	76.5
		Full	1.19	**77.2**

Table 6: Top block: WikiText (Generative); Bottom block: MNLI (Reasoning). We compare Memorization and Utility across MLP-only (Baseline), Attention-only, and Full adaptation. All configurations maintain effective unlearning (E-Hit <1.6%).

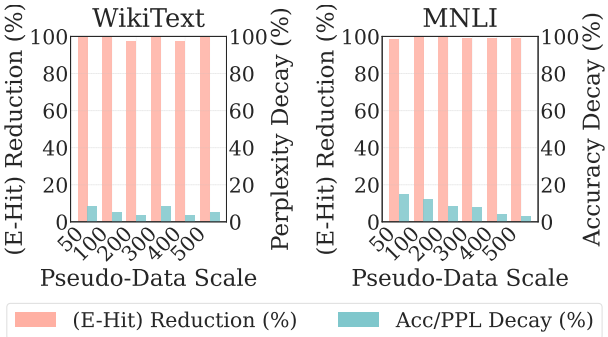

Figure 5: Blue bars (Utility Decay): Lower bars indicate better utility retention. Pink bars (Memory Reduction): Higher bars indicate better Privacy removal. Insight: Privacy reduction saturates rapidly (∼100 samples), while utility retention scales linearly with data volume, revealing a decoupling between the erasure of sparse privacy features and the preservation of dense semantic knowledge.

G.3 Pseudo-Data Scalability Evaluation

Scaling the size of the pseudo-dataset ($\hat{\mathcal{S}}$) from 50 to 500 samples reveals a functional asymmetry between privacy erasure and utility preservation (Fig. 5). Privacy metrics exhibit rapid convergence, with $|\hat{\mathcal{S}}| \approx 100$ achieving near-maximal leakage reduction. This suggests that the memorized entity manifold is intrinsically low-rank, allowing a compact surrogate set to effectively align with the privacy-related subspace and localize the unlearning critical threshold while minimizing intersection with the model's general functional space. In contrast, utility preservation—particularly for complex reasoning (MNLI)—scales with data density, requiring broader distributional coverage to provide sufficient contextual anchoring and prevent representational drift. Thus, while minimal data suffice for privacy guarantees, scaling $|\hat{\mathcal{S}}|$ is essential for safeguarding the integrity of the contextual manifold and preventing semantic over-correction.

H LoRA-based Unlearning Evaluation

We compare our selective unlearning method DFSU-su against two LoRA-based unlearning baselines on Pythia-1.4B, fixing Stages 1–2 outputs, the pseudo-PII dataset, and LoRA configurations (rank r=4, α_{lora}=32, MLP-only) to isolate the unlearning objective.

LoKU (Sun et al., 2025) applies an Inverted Hinge Loss (IHL) uniformly over all tokens in the forget data $\mathcal{S}$, with standard Gaussian LoRA initialization:

$$\mathcal{L}_{\text{IHL}} = \frac{1}{|T|}\sum_{t\in T}\max\big(0,\ \alpha_{\text{IHL}} + p_\theta(y_t) - \max_{y'\neq y_t} p_\theta(y')\big) \tag{21}$$

where T is the set of valid (non-padding) token positions, $p_\theta(y_t) = \text{softmax}(z_\theta)_{y_t}$ is the probability of ground-truth token y_t given logits z_θ, $\max_{y'\neq y_t} p_\theta(y')$ is the maximum probability among non-target tokens, and α_{IHL} is the hinge margin. The total objective adds cross-entropy ($\mathcal{L}_{\text{CE}}$) on retain data $\mathcal{D}_r$:

$$\mathcal{L}_{\text{LoKU}} = \mathcal{L}_{\text{IHL}}(\mathcal{S}) + \gamma \cdot \mathcal{L}_{\text{CE}}(\mathcal{D}_r) \tag{22}$$

where γ is the retain regularization strength. Following (Sun et al., 2025), we set α_{IHL}=1.0 and weight decay=0.01.

VILA (Lao et al., 2025) augments LoKU with Fisher Information-weighted LoRA initialization (WLRA). It estimates empirical Fisher Information $\hat{F}(W_0)$ (or its variance-corrected proxy $\text{Var}[\nabla\mathcal{L}]$ as per the VILA protocol) to compute the per-parameter importance ratio I_{W_0}:

$$I_{W_0} = \frac{\hat{F}_f(W_0)}{\hat{F}_r(W_0)+\epsilon},\quad \hat{F}(W_0) = \mathbb{E}_{x\sim\mathcal{D}}\big[(\nabla_{W_0}\mathcal{L}(x))^{\odot 2}\big] \tag{23}$$

where x is the input sequence, $W_0 \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ denotes the frozen pre-trained layer weights, $(\cdot)^{\odot 2}$ denotes element-wise squaring, ϵ is a stability constant, and the expectation is taken over all tokens in the respective dataset ($\mathcal{S}$ or $\mathcal{D}_r$). WLRA aligns the LoRA subspace with high-importance dimensions of the target matrix W_0:

$$\begin{aligned} U,\Sigma,V &= \text{SVD}_r\big(\text{diag}(\mathbf{d})\cdot W_0\big) \\ A &\leftarrow (V\sqrt{\Sigma})^\top,\quad B \leftarrow \text{diag}(\mathbf{d})^{-1}U\sqrt{\Sigma} \end{aligned} \tag{24}$$

where $\mathbf{d}_i = \sqrt{\sum_{j=1}^{d_{\text{in}}} I_{W_0}[i,j]}$ is the row-wise aggregated importance for output dimension i ($\mathbf{d}\in\mathbb{R}^{d_{\text{out}}}$), U,Σ,V are the truncated SVD components, A,B constitute the initialized LoRA adapter weights, and Σ absorbs the LoRA scaling factor α_{lora}/r. Initial parameters are drawn from $\mathcal{N}(0,\sigma^2_{\text{vila}})$ (σ_{vila}=0.05) before Fisher estimation, which is computed over 10 epochs with batch size 4. The resulting WLRA weights are scaled by s_{init}=0.1, followed by IHL optimization (Eq. 22). Each method uses its best-performing learning rate: 5×10^{-5} (MNLI) / 6×10^{-5} (WikiText) for DFSU-su, 10^{-4} for LoKU, and 2×10^{-4} for VILA. All other hyperparameters (effective batch size, epochs, warmup) are identical.

Scenario	Method	ERR	FRS	S-Exp	E-Hit	Utility
I: WikiText +PII (PPL ↓)	Original	21.40	43.70	32.20	24.76	7.02
	LoKU	0.20	21.01	7.60	3.32	7.21
	VILA	0.00	15.58	0.80	0.31	8.07
	DFSU-su	**0.00**	**4.42**	**3.00**	**1.00**	**7.23**
II: MNLI +PII (Acc% ↑)	Original	21.40	55.70	50.20	37.81	79.93
	LoKU	0.00	7.31	0.40	0.13	77.95
	VILA	0.20	18.06	7.00	2.88	79.24
	DFSU-su	**0.00**	**6.74**	**1.00**	**0.30**	**77.05**

Table 7: Comparison with LoRA-Based Unlearning Methods on Pythia-1.4B under the data-free setting. All methods use identical LoRA configurations (rank r=4, α_{lora}=32, MLP-only) and the same pseudo-PII dataset; only the unlearning objective and initialization differ. We report each method at its best privacy-utility operating point. LoKU (Sun et al., 2025) uses the Inverted Hinge Loss (IHL); VILA (Lao et al., 2025) further adds Fisher-weighted WLRA initialization atop IHL.

Comparison with LoRA-Based Unlearning Methods. As presented in Tab. 7, all methods mitigate privacy leakage relative to the original model, yet they exhibit divergent privacy-utility trade-offs. In the reasoning scenario (MNLI), although LoKU and VILA yield marginal accuracy improvements (+0.90 and +2.19 pp over DFSU-su), they incur substantial privacy degradation, with VILA retaining an FRS of 18.06% and E-Hit of 2.88%. This vulnerability is exacerbated in the generative scenario (WikiText), where both baselines demonstrate markedly incomplete privacy erasure, preserving FRS levels of 21.01% and 15.58%. Notably, LoKU fails to eliminate exact reconstruction (ERR $= 0.20\%$) and exposes $3.3\times$ more PII entities than our approach (E-Hit 3.32% vs. 1.00%). Conversely, DFSU-su consistently achieves superior erasure across both settings—reducing FRS to 4.42%, ERR to 0.00%, and E-Hit to 1.00%—while strictly preserving generative utility (PPL $= 7.23$).

LoRA Initialization Sensitivity. Previous studies establish that Fisher Information-weighted initialization (FILA (Sun et al., 2025), WLRA (Lao et al., 2025)) substantially accelerates sequence-level unlearning by aligning the LoRA subspace with forgetting-critical weight directions. However, this original protocol necessitates separate forget and retain datasets, rendering it inapplicable in our data-free setting. Furthermore, computing importance solely on the pseudo-PII dataset without retain normalization would align the subspace with the full-sequence next-token prediction gradient, fundamentally conflicting with the dual-stream optimization of DFSU-su. To rigorously evaluate whether advanced initialization benefits our method, we adapt FILA to a token-level formulation using the privacy mask $\mathbf{M}$:

$$
\begin{aligned}
I_{W_0}^{\text{token}} &= \frac{I_{\text{pii}}}{I_{\text{ctx}} + \epsilon} \\
I_{\text{pii}} &= \frac{1}{N_{\text{pii}}} \sum_{i,t:\, \mathbf{M}_{i,t}=1} (\nabla_{W_0} \ell_{i,t})^{\odot 2} \\
I_{\text{ctx}} &= \frac{1}{N_{\text{ctx}}} \sum_{i,t:\, \mathbf{M}_{i,t}=0} (\nabla_{W_0} \ell_{i,t})^{\odot 2}
\end{aligned}
\tag{25}
$$

where $\ell_{i,t}$ denotes the cross-entropy loss for token t in sequence i within a batch (expanding the notation from Eq. 21 to explicit sequence indices), $\mathbf{M}_{i,t} \in \{0, 1\}$ is the boolean mask value, and $N_{\text{pii}}, N_{\text{ctx}}$ are their respective token counts, which rigorously normalizes the extreme length imbalance between PII and context regions.

We evaluate this adaptation on Pythia-1.4B, computing Fisher over all $|\theta|$ parameters ($\sim$1.4B) to provide the most favorable statistical basis. Empirically, while standard Gaussian initialization proves fully sufficient for DFSU-su (achieving 77.05% MNLI accuracy and 7.23 WikiText PPL), token-level FILA-WLRA induces catastrophic model collapse *at initialization*. Prior to any optimization, the WLRA-initialized model degrades to near-random reasoning performance (MNLI 37.6%) and degenerated generation (WikiText PPL $> 8{,}000$). This collapse stems from an inherent statistical imbalance: PII tokens constitute merely $10-20\%$ of each sequence, yielding high-variance I_{pii} estimates. The consequently skewed $I_{\text{pii}}/I_{\text{ctx}}$ ratio is amplified by WLRA, producing adapter scaling factors ($> 40\times$) that acutely destabilize the model outside the valid training regime.

Crucially, even if this statistical skew were heuristically corrected, two fundamental incompatibilities persist. First, *gradient entanglement*: unlike original FILA computations on independent sequences, token-level Fisher estimates originate from the same forward pass, where self-attention inherently couples PII and context representations. Second, *objective mismatch*: WLRA statically aligns the rank-r (r=4) LoRA subspace with a single forgetting direction, whereas DFSU-su demands simultaneous gradient ascent ($\mathcal{L}_{\text{priv}}$) and descent ($\mathcal{L}_{\text{gen}}$) within that identical subspace.

While Fisher-weighted pre-computation is highly effective for standard sequence-level unlearning, the aforementioned incompatibilities

make it challenging to directly adapt to our token-level scenario. Nevertheless, DFSU-su's dual-stream objective already provides an *implicit* importance weighting at every training step: $\mathcal{L}_{\text{priv}}$ naturally concentrates gradients on PII-relevant weight directions while $\mathcal{L}_{\text{gen}}$ simultaneously anchors context-relevant directions. By dynamically refining these directions during training, our method inherently achieves the functional effect that FILA attempts to encode statically at initialization, explaining why standard Gaussian initialization is fully sufficient for DFSU-su.