# 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 using Pythia models 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.

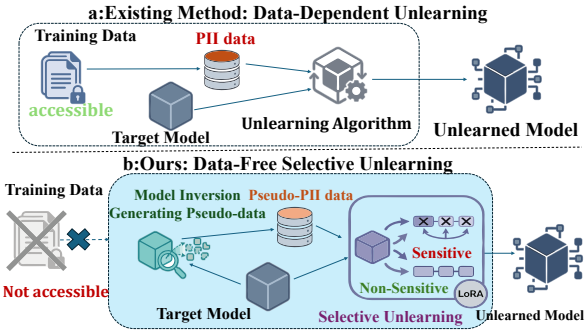


Figure 1: A comparison of (a) data-dependent unlearning and (b) data-free selective unlearning.

Motivated by the cognitive phenomenon that specific memories can be attenuated by suppressing internal representations without re-exposure to sensitive contents, we study *data-free selective unlearning* (Fig. 1b): removing memorized PII from a pre-trained LLM *post hoc*, using only model parameters and without accessing 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 explicit data supervision, the optimization signal is weakly constrained, and naive updates diffuse across entangled representations, leading to either incomplete privacy removal or unnecessary utility degradation.

A key practical observation is that defenders often know the *type* of information to be forgotten (e.g., IP addresses, device identifiers) even when the exact training instances are unavailable. We leverage this prior as a directional cue and propose to synthesize an effective surrogate "forget set" via model inversion, repurposing inversion attacks as a defensive tool. Building on this insight, we introduce DFSU, a data-free privacy-preserving framework that removes sensitive PII from an LLM 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 a target LLM; (ii) we generate pseudo-PII samples 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 suppress identified sensitive tokens while anchoring surrounding context to preserve utility.

We evaluate DFSU on both generative (WikiText-103) (Merity et al., 2017a) and reasoning/classification (MNLI) (Williams et al., 2018a) tasks using pretrained Pythia models (160M/410M/1.4B) (Biderman et al., 2023) and sensitive data from the AI4Privacy dataset (AI4Privacy, 2024). Across scales 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 can 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 post-hoc unlearning methods are intrinsically data-dependent, requiring access to ground-truth sensitive examples. Gradient-ascent (GA) approaches (Jang et al., 2023) maximize loss on private samples but can induce catastrophic collapse, degrading general language competence alongside the targeted facts (Yuan et al., 2025; Xing et al., 2025). Negative Preference Optimization (NPO) (Zhang et al., 2024) mitigates instability by anchoring updates to a reference distribution, yet still assumes a precisely specified forget set. Related model-editing work frames unlearning as localized knowledge suppression: for instance, Private Memorization Editing (PME) (Ruzzetti et al., 2025) first detects memorized PII via extraction and then edits feed-forward layers to reduce its emission. 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. By design, 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, aiming to reconstruct training inputs from model representations or outputs. Methods such as Vec2Text (Mor-

ris 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., 2025b), we propose a paradigm shift by leveraging inversion for defensive purposes. By treating a model's own memorized logits as a generative prior, our DFSU synthesizes privacy-relevant pseudo-samples to bridge the data-free gap. Crucially, instead of retraining on noisy synthetic data, we integrate inversion with token-level selective masking to suppress target PII without requiring access to the original training data.

## 3 Methodology

### 3.1 Problem Formulation

Let $\mathcal{M}_\theta$ be an LLM parameterized by $\theta$, which has inadvertently memorized a sensitive set $\mathcal{S}$ within its training set $\mathcal{D}$, (i.e., $\mathcal{S} \subset \mathcal{D}$). In the standard unlearning setting, one seeks to update $\theta \rightarrow \theta^*$ 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, existing unlearning algorithms like GA inherently require access to $\mathcal{S}$ to compute the forgetting gradient:

$$\nabla_\theta \mathcal{L}_{\text{forget}} = -\nabla_\theta \mathbb{E}_{s\in\mathcal{S}}\left[\log P_\theta(s)\right] \tag{2}$$

In our data-free setting, $\mathcal{S}$ is unavailable. In fact, model parameters $\theta$ act as a holographic storage of $\mathcal{S}$. We aim to find a function mapping model parameters $\theta$ to the sensitive set $\mathcal{S}$. Formally:

$$\mathcal{D}_{\text{pseudo}} = \mathcal{I}_\phi(\theta) \approx_{\text{semantic}} \mathcal{S} \tag{3}$$

where $\approx_{\text{semantic}}$ denotes semantic approximation.

**Problem.** We substitute the unavailable ground truth $\mathcal{S}$ with model-derived surrogates $\mathcal{D}_{\text{pseudo}}$ in the forgetting objective. Specifically, we update the model parameters by minimizing:

$$\begin{aligned} \theta^*_{\text{pseudo}} &= \arg\min_\theta \ \mathbb{E}_{\hat{s}\in\mathcal{D}_{\text{pseudo}}}\left[\log P_\theta(\hat{s})\right] \\ \text{subject to} \quad & \mathcal{L}(\theta; \mathcal{D}\setminus\mathcal{S}) \approx \mathcal{L}(\theta_0; \mathcal{D}\setminus\mathcal{S}) \end{aligned} \tag{4}$$

where $\hat{s}$ denotes a pseudo-sample from $\mathcal{D}_{\text{pseudo}}$, and $\theta^*_{\text{pseudo}}$ represents the parameters obtained via surrogate-based unlearning. Note that this formulation mirrors Eq. 1, with the critical distinction that $\mathcal{S}$ is replaced by its inversion-derived approximation $\mathcal{D}_{\text{pseudo}}$.

**DFSU.** To address the problem in Eq. 4, we propose a novel three-stage framework, DFSU, as illustrated in Fig 2. This framework effectively synthesize the sensitive set $\mathcal{S}$'s surrogates $\mathcal{D}_{\text{pseudo}}$ and then performs the unlearning process. Specifically, the pipeline consists of: **(1) Inversion Model Training**, which trains a logit-based inversion model to capture memorized PII patterns from the target LLM; **(2) Pseudo-Data Synthesis and Annotation**, where we query the target model with entity-swapped candidates, employ the trained inverter to synthesis pseudo-PII $\mathcal{D}_{\text{pseudo}}$, and annotate $\mathcal{D}_{\text{pseudo}}$ via few-shot prompting; and **(3) Selective Unlearning**, which leverages a dual-stream contrastive objective to maximize the loss on sensitive tokens while preserving non-sensitive contexts under a LoRA-constrained optimization. We present our algorithm in Appendix 7.

### 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 $M_{\text{target}}$, we train an inverter model $M_{\text{inv}}$ (a sequence-to-sequence transformer) to recover the input text $\mathbf{x}$ from $M_{\text{target}}$'s log-probability distribution $P_t$ at the final token position.

**Inverter Training.** We train an inverter $M_{\text{inv}}$ to reconstruct texts from the target model $M_{\text{target}}$'s log-probabilities $P_t$. The inverter maps $P_t$ to its vocabulary via token matching, computes soft embeddings as weighted sums of its word embeddings using $P_t$ as weights, and applies a learnable projection $\phi$ before decoding. We minimize the standard sequence-to-sequence cross-entropy loss $\mathcal{L}_{\text{inv}}$ on pairs $(\mathbf{x}, P_t)$ of original texts and pre-computed probabilities. High-quality inversion (F1 $> 30\%$, BLEU $> 15\%$) enables generation of pseudo-labels approximating the target model's training distribution for our selective unlearning framework.

### 3.3 Pseudo-PII Synthesis and Annotation

After training the inversion model, we synthesize and annotate pseudo-PII using a pipeline as shown Fig 2 (Step 2). Firstly, we reuse the syntactic templates from the PII data that used to train the target

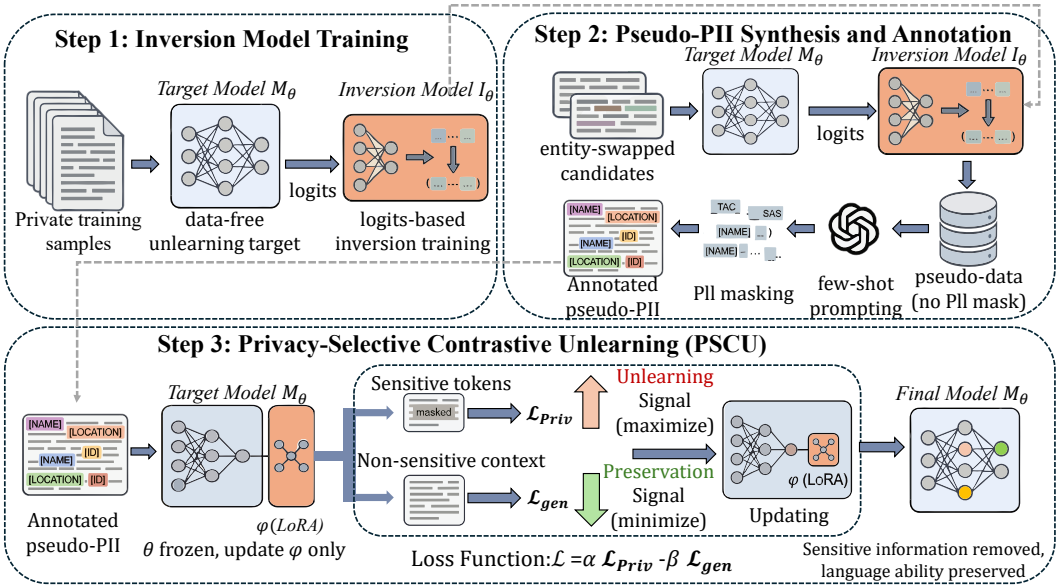


Figure 2: An overview of our DFSU framework.

model and replace all sensitive entities with random substitutes drawn from a public, disjoint pool. Secondly, we query the target model using entity-swapped candidates to extract internal confidence distributions (logits) which harbor memorized PII. Third, our trained inverter $\mathcal{I}_\phi$ in Sec. 3.2 decodes these logits into pseudo-PII sequences, which approximate the target model's training data distribution. Fourth, we annotate these decoded pseudo-PII sequences using few-shot prompting. Specifically. we provide examples containing start and end position for privacy-sensitive entities to an LLM and prompt it to mark locations of privacy-sensitive entities in the generated pseudo-PII, thereby generating annotated pseudo-PII.

### 3.4 Privacy-Selective Contrastive Unlearning

Given the surrogate dataset $\mathcal{D}_{\text{pseudo}}$, we introduce PSCU to ensure selective forgetting via a constrained update space and a token-localized objective. Concretely, we freeze the pre-trained weights $\theta$ and optimize only LoRA parameters $\phi$, 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:

$$\mathcal{L}_{\text{priv}} = \frac{\sum \mathbf{M}_{i,t} \cdot \ell_{i,t}}{\sum \mathbf{M}_{i,t} + \epsilon} \tag{5}$$

$$\mathcal{L}_{\text{gen}} = \frac{\sum (1 - \mathbf{M}_{i,t}) \cdot \ell_{i,t}}{\sum (1 - \mathbf{M}_{i,t}) + \epsilon} \tag{6}$$

We then minimize the following contrastive objective:

$$\mathcal{J}(\phi) = \alpha \cdot \mathcal{L}_{\text{gen}} - \beta \cdot \mathcal{L}_{\text{priv}} \tag{7}$$

where $\alpha$ and $\beta$ are hyperparameters balancing preservation and erasure.

## 4 Experimental Setup

**Datasets.** We construct our dataset by injecting sensitive privacy data into a general language corpus. We employ the AI4Privacy PII dataset (AI4Privacy, 2024) as the source of sensitive privacy data, blending it with two established general corpora: WikiText-103 (Merity et al., 2017b) for generative tasks and the MNLI corpus (Williams et al., 2018b) for classification tasks. To study memorization, we partition 500 unique PII samples into 10 disjoint groups of 50 samples each. For group $G_i$, we construct a scaled dataset by replicating (augmenting) each sample exactly $10i$ times, yielding exposure levels from 10 to 100 repetitions (Li et al., 2024). Crucially, our data-free unlearning algorithm, DFSU, never accesses the injected samples; instead, it queries the model via entity swapping to ensure strict non-reproducibility.

**Metrics.** We evaluate the performance of DFSU along two dimensions: the preservation of general model utility, and the effectiveness of privacy protection via unlearning. In terms of model utility, we report standard performance metrics: perplexity (PPL) for generative tasks and accuracy (Acc) for classification tasks. In terms of unlearning effectiveness, we employ Exact Reconstruction Rate (ERR) and Fractional Recon-

struction 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. More details of these metrics are presented in Appendix 7

**Implementation Details.** We evaluate Pythia (160M/410M/1.4B). Each model is fully fine-tuned via continued pre-training on the injected corpus for 6 epochs (AdamW, cosine schedule, bf16; peak lr 2–6 $\times 10^{-5}$ depending on scale). We use a single inverter $\mathcal{I}_\phi$ (Flan-T5-Large) trained only on Pythia-410M, and reuse it across all Pythia scales based on their shared architecture. Training runs for 30 epochs (bs=256, lr=$5\times10^{-4}$), keeping embeddings in FP32 for numerical stability. For unlearning, we apply PSCU with LoRA on MLP modules (rank $r = 4$, $\alpha_{\text{lora}} = 32$, dropout 0). We set the dual-objective weights to $\lambda_1 = \lambda_2 = 1.0$ and optimize for 10 epochs with AdamW (effective bs=16; lr $5 \times 10^{-5}$–$10^{-4}$).

**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 PSCU 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

**Evaluation Protocol.** We evaluate DFSU in two tiers to separate *mechanistic validity* from *deployment realism*. **(i) Injection-Based Simulation:** (Sec. 5.1) we use an injection-based protocol where PII is inserted into a known corpus, and evaluate unlearning under two task regimes: **Scenario I** (WikiText+PII) emphasizing generative language modeling, and **Scenario II** (MNLI+PII) emphasizing NLU-style reasoning. **(ii) In the Wild Evaluation:** (Sec. 5.2) we apply DFSU to an *unaltered, production-ready checkpoint* (no artificial injection and no access to the original pre-training data), and measure behavioral shifts on PII-related prompts. We further substantiate PSCU with targeted ablations (Sec. 5.3) and hyperparameter robustness analyses (Sec. 5.4), focusing on the selective masking mechanism and its stability under different LoRA parameterizations. The results consistently indicate that PSCU admits a reliable regime that preserves utility while delivering thorough privacy erasure.

## 5.1 Performance Improvement of DFSU.

We now interpret Tab. 1 following the two controlled scenarios. table 1 summarizes results across three model scales (Pythia (160M/410M/1.4B)) under the controlled protocol, reporting privacy leakage via ERR, FRS, S-Exp, and E-Hit (lower is better), and utility via PPL (WikiText) or Accuracy (MNLI).

**Scenario I: WikiText+PII (Generative).** We test whether privacy unlearning can suppress memorization while preserving language modeling utility. All three original checkpoints 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 PSCU toward oracle-level privacy suppression with limited degradation of generative utility.

**Scenario II: MNLI+PII (Reasoning).** We next examine whether unlearning preserves NLU capability under high initial leakage. Original models again show severe privacy risk (e.g., S-Exp 50.20% for Pythia-1.4B), whereas DFSU drives 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 PSCU settings).

## 5.2 In the Wild Evaluation

While injection-based simulations validate DFSU under a *known* memorization profile, real-world remediation must operate on *unaltered* production checkpoints whose privacy leakage is *unknown* a priori and whose original pre-training data is unavailable. To assess this setting, we apply DFSU

| Scenario I: WikiText+PII (Generative) | | | | | | |
|---|---|---|---|---|---|---|
| Model | Method | Privacy Metrics (↓) | | | | Performance |
| | | ERR (%) | FRS (%) | S-Exp (%) | E-Hit (%) | PPL (↓) |
| Pythia-160M | Original Model | 0.80 | 19.75 | 9.80 | 4.64 | 13.71 |
| | Original Data (Oracle) | 0.00 | 11.68 | 2.20 | 0.69 | 14.11 |
| | **DFSU (Ours)** | **0.00** | **13.38** | **2.40** | **0.75** | **14.09** |
| Pythia-410M | Original Model | 20.40 | 46.50 | 36.80 | 28.78 | 8.39 |
| | Original Data (Oracle) | 0.00 | 3.46 | 0.20 | 0.06 | 8.69 |
| | **DFSU (Ours)** | **0.00** | **3.88** | **0.40** | **0.13** | **8.83** |
| Pythia-1.4B | Original Model | 21.40 | 43.70 | 32.20 | 24.76 | 7.02 |
| | Original Data (Oracle) | 0.00 | 5.83 | 2.00 | 0.63 | 7.13 |
| | **DFSU (Ours)** | **0.00** | **4.42** | **3.00** | **1.00** | **7.23** |

| Scenario II: MNLI+PII (Reasoning) | | | | | | |
|---|---|---|---|---|---|---|
| Model | Method | Privacy Metrics (↓) | | | | Performance |
| | | ERR (%) | FRS (%) | S-Exp (%) | E-Hit (%) | Acc (%) (↑) |
| Pythia-160M | Original Model | 17.00 | 47.99 | 38.60 | 30.85 | 45.28 |
| | Original Data (Oracle) | 0.00 | 6.51 | 0.80 | 0.25 | 44.06 |
| | **DFSU (Ours)** | **0.00** | **8.99** | **0.40** | **0.13** | **43.38** |
| Pythia-410M | Original Model | 17.60 | 53.05 | 45.20 | 34.73 | 70.44 |
| | Original Data (Oracle) | 0.00 | 11.10 | 1.80 | 0.63 | 69.90 |
| | **DFSU (Ours)** | **0.00** | **11.85** | **1.00** | **0.38** | **68.45** |
| Pythia-1.4b | Original Model | 21.40 | 55.70 | 50.20 | 37.81 | 79.93 |
| | Original Data (Oracle) | 0.00 | 6.42 | 1.80 | 0.56 | 77.21 |
| | **DFSU (Ours)** | **0.00** | **7.11** | **1.20** | **0.38** | **77.05** |

Table 1: **Results for Injection-Based Simulation.** We report privacy leakage by ERR/FRS/S-Exp/E-Hit (↓) and utility by WikiText perplexity (PPL) or MNLI accuracy. **Original Data (Oracle)** employs PSCU to perform machine unlearning using ground-truth PII targets; **Data-Free (Ours)** uses inversion-derived surrogates. Across both scenarios, **Data-Free** attains *zero ERR* at all scales and remains close to the oracle in both privacy and utility.

directly to the **original Pythia-1.4B checkpoint** (i.e., without any artificial PII injection), and use the same 500-sample AI4Privacy corpus to synthesize inversion-based surrogate targets. To assess post-hoc changes in generation behavior, we evaluate the model on 100 low-perplexity PII-related prompts and use greedy decoding to eliminate sampling variance. Tab. 2 reports representative outputs. Compared to the original checkpoint, the DFSU-unlearned model tends to substitute PII-like entities with alternative yet contextually plausible realizations, while largely preserving grammaticality and topical coherence. Overall, these examples are *consistent with a shift in the conditional distribution over PII-like entity realizations*, rather than a narrow removal of a single memorized verbatim suffix.

### 5.3 Ablation Study

**PSCU Outperforms GA.** Our **P**rivacy-**S**elective **C**ontrastive **U**nlearning (**PSCU**) provides a principled alternative to full-sequence **G**radient **A**scent (**GA**) for privacy removal under parameter-efficient updates. In a controlled ablation—holding all hyperparameters, LoRA target modules, and training budgets constant and varying only the *locus of loss maximization*—we find that indiscriminate full-sequence ascent is brittle. On WikiText (Fig 3, pushing GA to match PSCU's near-zero leakage regime on Pythia-410M (E-Hit $\approx 0.13\%$) causes perplexity to explode beyond $4 \times 10^4$, whereas PSCU attains comparable privacy reduction with stable PPL $= 8.83$ (oracle-comparable). A similar pattern holds for reasoning on MNLI (Fig 3: GA can drive E-Hit to $0.0\%$ but at an unacceptable utility cost (accuracy drops to $57.35\%$), while PSCU achieves near-identical privacy (E-Hit $= 0.38\%$) with substantially higher accuracy ($68.45\%$), yielding a Pareto-superior operating point. Overall, these results indicate that effective unlearning depends less on the magnitude of updates than on their *directionality and localization*: by confining ascent to sensitive entity tokens and anchoring the surrounding context, PSCU selectively removes privacy signals while avoiding the collateral degradation induced by sequence-wide GA.

| Prompt (Prefix) | Original Model | DFSU-Unlearned |
|---|---|---|
| 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. |

Table 2: Representative greedy-decoded suffixes from the original and DFSU-unlearned **Pythia-1.4B** model on PII-related prefixes. Highlighted spans illustrate how entity-level realizations shift post-unlearning while contextual coherence is preserved.

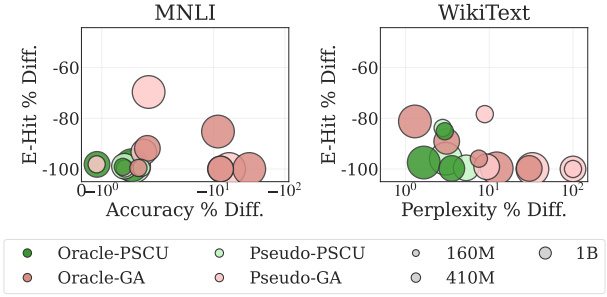


Figure 3: **Ablation Analysis across Models and Scenarios: PSCU (Ours) vs. Full-Sequence Gradient Ascent (GA).** Left figure shows results on MNLI (Accuracy drop), and right figure on WikiText (Perplexity increase). Circle size correspond to model sizes: 160M, 410M, and 1.4B. **Observation:** Across all scales, our selective PSCU method (Green circles) consistently achieves better privacy-utility trade-offs (bottom-left region) compared to the full-sequence GA baseline (Pink circles), which suffers from severe utility degradation to achieve comparable unlearning efficacy.

**Uniform Privacy *vs.* Task-Specific Utility.** Prior work suggests that the choice of LoRA target modules (e.g., MLP-only vs. Attention-only) can materially affect the privacy–utility trade-off in parameter-efficient unlearning (Chen et al., 2025a). To assess whether our PSCU depends on a particular parameter subspace, we perform a controlled comparison of three LoRA configurations—**MLP-only** (feed-forward, default), **Attention-only** (QKV+Dense), and **Full**. The results reveal a clear dichotomy: *privacy is largely architecture-agnostic, whereas utility is task- and module-dependent*. Across model scales and both tasks, all configurations achieve deep unlearning with consistently low leakage (E-Hit $<$ 1.6%, and as low as 0.00% on Pythia-410M with Attention/Full), indicating a strong uniform privacy property. This invariance supports our central hypothesis that, once the forgetting signal is precisely localized to sensitive entity tokens, its quality dominates the optimization dynamics, making the specific LoRA module choice secondary for privacy erasure. In contrast, utility preservation exhibits distinct modular sensitivity: for generation (WikiText), **MLP-only** is consistently more stable (e.g., on Pythia-410M, **Full** increases PPL from 8.83 to 10.23), suggesting that broader adaptation injects excess plasticity and drifts away from the pre-trained manifold, whereas restricting updates to MLPs yields a more controlled intervention; for reasoning (MNLI), Attention-based adaptation can be competitive (e.g., highest accuracy 69.9% on 410M), consistent with attention pathways contributing to logical coherence. Taken together, these findings motivate **MLP-only LoRA** as a robust default that preserves *uniform privacy* while offering a Pareto-efficient balance between computational cost and task-specific utility.

**Early Privacy Saturation, Late Utility Recovery.** We ablate the pseudo-dataset scale (50–400 samples) to quantify the data requirement for privacy erasure. Fig 5 shows a clear *asymmetry* between privacy and utility: **privacy saturates early**, with near-maximal leakage reduction already achieved at 100 pseudo-samples, matching the 500-sample baseline in both WikiText and MNLI. This suggests the forgetting signal for memorized entities is redundant and effectively low-dimensional. In contrast, **utility scales with data and task complexity**: generation is relatively robust to scarcity

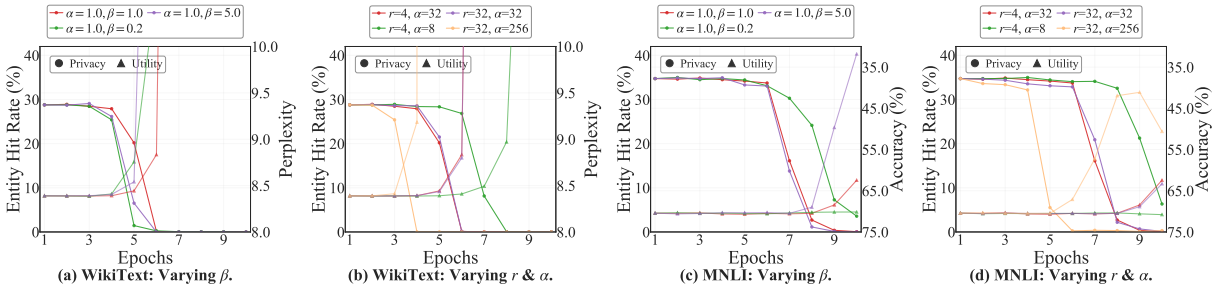


Figure 4: Privacy-utility trajectories under varying privacy weight $\beta$ and LoRA configurations across WikiText and MNLI scenarios.

| | **Memorization** (E-Hit(%) ↓) | | | **Utility** (PPL ↓) | | |
|---|---|---|---|---|---|---|
| **Model** | **MLP** | **Attn** | **Full** | **MLP** | **Attn** | **Full** |
| 160m | 0.75 | 1.57 | 0.63 | 14.09 | 14.44 | 15.58 |
| 410m | 0.13 | 0.00 | 0.00 | 8.83 | 9.98 | 10.23 |
| 1.4b | 1.00 | 0.88 | 1.38 | 7.23 | 7.29 | 7.17 |
| | **Memorization** (E-Hit(%) ↓) | | | **Utility** (Acc(%) ↑) | | |
| **Model** | **MLP** | **Attn** | **Full** | **MLP** | **Attn** | **Full** |
| 160m | 0.12 | 0.13 | 0.00 | 43.4 | 46.4 | 45.2 |
| 410m | 0.38 | 0.31 | 0.13 | 68.5 | 69.9 | 67.0 |
| 1.4b | 0.38 | 1.19 | 1.19 | 77.1 | 76.5 | 77.2 |

Table 3: **LoRA Target Module Robustness.** 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%).

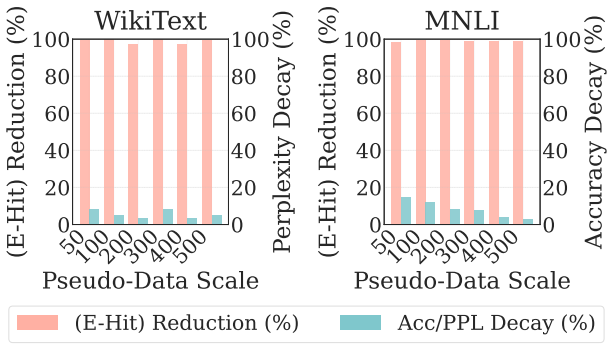


Figure 5: **Benchmarking Data Efficiency.** The results from 500 samples represent the test results of the complete dataset. **Blue bars (Utility Decay):** Lower bars indicate better utility retention. **Pink bars (Memory Reduction):** Higher bars indicate better Privacy removal. **Insight:** Privacy 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.

(WikiText ∼5.4% decay at Scale 100), whereas reasoning requires denser support to avoid semantic over-erasure (MNLI ∼12.1% decay at Scale 100 vs. ∼2.8% at full scale). Practically, small pseudo-sets suffice for strong privacy guarantees, while larger scales mainly improve reasoning fidelity via better distributional coverage.

### 5.4 Hyperparameter Study: $\beta$, $r$, and $\alpha$.

To locate hyperparameter regimes that deliver *complete* privacy mitigation (E-Hit $< 1\%$) with minimal utility loss, we sweep the privacy weight $\beta$ and the LoRA parameterization (rank $r$, scaling $\alpha$) for Pythia-450M unlearning on pseudo-data over Epochs 1–10, tracking E-Hit, WikiText PPL, and MNLI accuracy (Fig 4). Two constraints emerge. **(1) $\beta$ controls completeness vs. stability:** $\beta = 1.0$ offers the best operating point, reaching E-Hit $< 1\%$ by Epoch 6 with $< 6\%$ PPL increase; underweighting ($\beta = 0.2$) yields under-erasure (E-Hit $> 1\%$), while over-weighting ($\beta = 5.0$) speeds forgetting but destabilizes optimization (rapid PPL blow-up), making early stopping essential. **(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 higher-capacity settings such as $(r = 32, \alpha = 32)$ or $(r = 32, \alpha = 256)$ introduce delayed or immediate 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.

## 6 Conclusion

We propose a novel three-stage framework, DFSU, to address the data-free privacy-preserving for LLMs. Extensive experiments on the AI4Privacy dataset using Pythia models demonstrate that our method achieve a privacy-utility trade-off competitive with Oracle-based unlearning.

## 7 Limitations

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.

Xinyu Chen and 1 others. 2025a. Towards robust and parameter-efficient knowledge unlearning for llms. *arXiv preprint arXiv:2502.01876*.

Yiyi Chen, Qiongkai Xu, and Johannes Bjerva. 2025b. 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)*.

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*.

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.

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.

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.

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.

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.

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)*.

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. Algorithm of DFSU

We present our algorithm as follows:

**Algorithm 1** Data-Free Selective Unlearning (DFSU)

**Input:** Target Model $\mathcal{M}_\theta$, Training Corpus $\mathcal{D}_{\text{train}}$, Entity-Swapped Candidates $\mathcal{C}$, Hyperparameters $\alpha, \beta, \eta$

**Output:** Unlearned Model $\mathcal{M}_{\theta^*}$

**Stage 1: Inversion Model Training**
Pre-compute logits: $\mathcal{D}_{\text{logits}} \leftarrow \{(\mathcal{M}_\theta(x), x) \mid x \in \mathcal{D}_{\text{train}}\}$
Train inverter $\mathcal{I}_\phi$ via: $\min_\phi \mathbb{E}_{(L,X)\sim\mathcal{D}_{\text{logits}}}[-\log P_\phi(\mathcal{I}_\phi(L) = X)]$

**Stage 2: Pseudo-PII Synthesis and Annotation**
Initialize $\mathcal{D}_{\text{pseudo}} \leftarrow \emptyset$
**for** each candidate $c \in \mathcal{C}$ **do**
  $\hat{x} \leftarrow \mathcal{I}_\phi(\mathcal{M}_\theta(c))$ {Decode logits to pseudo-text}
  $\mathbf{M} \leftarrow$ PromptLLM("Mark PII in: " $\oplus\ \hat{x}$) {Few-shot annotation}
  $\mathcal{D}_{\text{pseudo}} \leftarrow \mathcal{D}_{\text{pseudo}} \cup \{(\hat{x}, \mathbf{M})\}$
**end for**

**Stage 3: Privacy-Selective Contrastive Unlearning**
Freeze $\theta$; Initialize LoRA adapter $\phi \leftarrow \phi_0$
**for** step $t = 1, \ldots, T$ **do**
  Sample $(\mathbf{X}, \mathbf{M}) \sim \mathcal{D}_{\text{pseudo}}$
  Compute token-wise loss: $\ell(\mathbf{X}) \leftarrow -\log P_{\theta,\phi}(\mathbf{X})$
  $\mathcal{L}_{\text{priv}} \leftarrow \frac{\mathbf{M}^\top \ell(\mathbf{X})}{\|\mathbf{M}\|_1}$, $\mathcal{L}_{\text{gen}} \leftarrow \frac{(1-\mathbf{M})^\top \ell(\mathbf{X})}{\|1-\mathbf{M}\|_1}$
  $\phi \leftarrow \phi - \eta\nabla_\phi(\alpha\mathcal{L}_{\text{gen}} - \beta\mathcal{L}_{\text{priv}})$
**end for**

**Return** $\mathcal{M}_{\theta^*} \leftarrow \mathcal{M}_\theta \oplus \phi$

## B. Evaluation 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:

$$\text{ERR} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbb{I}\left(\hat{s}_i^{(j)} = s_i\right). \tag{8}$$

$$\text{FRS} = 1 - \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\frac{\text{Lev}(s_i, \hat{s}_i^{(j)})}{\max\left(|s_i|, |\hat{s}_i^{(j)}|, 1\right)}. \tag{9}$$

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 $\tau$ 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:

$$\text{S-Exp} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\exists j, \exists e \in E_i :\ e \subseteq \hat{s}_i^{(j)}\right]. \tag{10}$$

$$\text{E-Hit} = \frac{\sum_{i=1}^{N}\left|\left\{e \in E_i \mid \exists j :\ e \subseteq \hat{s}_i^{(j)}\right\}\right|}{\sum_{i=1}^{N}|E_i|}. \tag{11}$$

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 $e$ ap-

pears 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.