TRACE: Data-Free Text Reconstruction Attacks against Approximate Unlearning in LLMs

Mengying Zhang[1] **Derui Wang**[2] **Ibrahim Khalil**[1] **Bowen Liu**[4]
Xiaoyu Xia[1*] **Minhui Xue**[2,3]
[1]RMIT University [2]CSIRO [3]Adelaide University
[4]The Hong Kong University of Science and Technology

Abstract

As large language models (LLMs) face growing demands for data removal driven by regulatory compliance, copyright concerns, and privacy protection, approximate unlearning has emerged as a practical alternative to expensive full retraining. In practice, pre- and post-unlearning model snapshots are often maintained for auditing and rollback, introducing a previously overlooked side channel. This paper demonstrates that approximate unlearning in LLM can leave residual traces within model snapshots, which may be recoverable and thus enable the privacy breaches that approximate unlearning is intended to prevent. Here, we propose TRACE, a framework that reconstructs unlearned training data given access to pre- and post-unlearning model snapshots, without any data-specific prior knowledge. To achieve this, TRACE identifies a compact set of candidate tokens from embedding-level parameter differences to constrain the search space, and assembles sequences via a contrastive scoring mechanism based on the perplexity gap between the two models. A two-stage exploration-then-completion strategy enables the recovery of multiple distinct samples. Extensive experiments on various datasets, including TOFU, MUSE, and WMDP, across multiple LLMs and six representative unlearning methods, demonstrate that TRACE achieves high-fidelity reconstruction. This paper exposes significant privacy risks in current LLM approximate unlearning deployments and highlights the need for defenses against parameter-difference leakage and appropriate management of model snapshot access. The code is publicly available at `https://github.com/mengyingz969/TRACE`.

1 Introduction

The rapid adoption of large language models (LLMs) has intensified the need for effective data removal mechanisms, driven by regulatory requirements such as the EU's GDPR [1] and AI Act [2], copyright compliance [3, 4, 5], and privacy [6, 7]. Machine unlearning [8, 9, 10] aims to remove the influence of a specified subset of training data from a trained model. In practice, exact machine unlearning [8], retraining models without the target data, remains computationally prohibitive for modern LLMs. As a result, mainstream real-world deployments rely on approximate unlearning methods [11, 12, 13, 14, 15, 16, 17], which aim to forget specific data through targeted parameter updates without retraining. To support auditing, rollback, and safety verification, it is common to retain multiple model snapshots from before and after unlearning.

This raises a critical question: *given access to original and unlearned models, to what extent do current unlearning methods leak sensitive information about the data they are intended to remove?*

Existing work [18] shows that differences between the original and unlearned model parameters can be used to construct surrogate gradient signals, which are then exploited through gradient-

*Corresponding author: Xiaoyu Xia.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

matching inversion to reconstruct removed images. However, this cannot be extended to generative LLMs, as the reconstruction optimization struggles to converge and often produces uninterpretable or disjointed outputs without the "anchor" provided by labels. More recently, efforts have been made to investigate the extraction of unlearned data from LLMs under the assumption that partial prefixes of the target texts are available, demonstrating that the remaining content can be recovered by exploiting discrepancies in output likelihoods between pre- and post-unlearning models [19].

However, this reliance on prefix priors leaves an important open question: *is unlearned data in LLMs secure without any data-specific prior knowledge?*

We answer this question by introducing TRACE, a framework for reconstructing text that has been intentionally unlearned by tracing latent parameter footprints between pre- and post-unlearning snapshots, without requiring any data-specific prior knowledge. First, we identify that parameter differences in the embedding layer between pre- and post-unlearning models expose the tokens involved in the unlearning process, effectively reducing the reconstruction task from open-ended generation over the full vocabulary to assembling sequences from a compact set of candidate tokens. Second, to resolve the combinatorial complexity of sequence reassembly, we develop a contrastive search strategy driven by a novel contrastive fluency score (CFS). This scoring mechanism integrates unlearning-induced divergence with linguistic plausibility, effectively navigating the search toward trajectories corresponding to unlearned sequences. Third, to enable reconstruction of multiple unlearned samples, we design a two-stage "exploration-then-completion" architecture. Stage I performs broad exploration to capture diverse prefix candidates, while Stage II completes them into full sequences. Our contributions are summarized as follows:

- We reveal that approximate unlearning leaves a latent footprint in parameters, enabling reconstruction of text that has been intentionally unlearned, without any data-specific prior knowledge.
- We propose a framework named TRACE that achieves the reconstruction of multiple distinct unlearned samples without any data-specific prior knowledge, by utilizing parameter differences to effectively constrain the vast search space and employing a contrastive search strategy within a two-stage architecture.
- We conduct comprehensive experiments on various benchmarks including TOFU, MUSE, and WMDP, across multiple LLMs and six representative unlearning methods, showing that TRACE consistently achieves high-fidelity reconstruction.

2 Related Work

Machine Unlearning in LLMs. Machine unlearning [8] aims to remove the influence of specific training samples from a trained model. Exact unlearning, which retrains the model on the remaining data, is widely regarded as the gold standard for data removal. However, exact unlearning is impractical for large language models (LLMs) due to their massive training scale and prohibitive computational cost. As a result, most existing work focuses on approximate unlearning via efficient parameter updates. Gradient ascent (GA) [11, 12, 13, 14] forgets samples by reversing learning on target data, i.e., applying updates that increase the loss on the data to be removed. Gradient difference (GD) [15] balances forgetting and utility by maximizing loss on forget data while minimizing loss on retain data. KL-regularized methods (KL) [16] introduce Kullback–Leibler divergence regularization on a retained set to preserve the model's behavior on retained data during unlearning. Negative Preference Optimization (NPO) [17] reformulates unlearning as preference optimization with a reference model to stabilize training. Suppression-intended unlearning [20, 21, 22, 23, 24, 25] suppresses sensitive outputs rather than removing knowledge and is beyond our scope.

Training Data Reconstruction. Recent studies show that training data can be reconstructed from models. Gradient-based approaches recover samples from leaked gradients [26, 27, 28, 29]. To name a few, LAMP [26] reconstructs text using a language model prior and alternating optimization, while Gupta et al. [27] use beam search with gradient-guided reordering. Petrov et al. [28] exploit low-rank self-attention gradients for exact inversion, and Xie et al. [29] demonstrate reconstruction under parameter-efficient fine-tuning. Beyond gradients, reconstruction from internal representations has also been explored, such as embedding leakage [30] and activation inversion attacks [31]. However, these attacks rely on access to gradients or intermediate activations during training, and thus cannot be directly applied when only pre- and post-unlearning model snapshots are available.

Unlearning Inversion Attacks. Unlearning inversion attacks were first introduced in the context of image classification to reconstruct unlearned images by approximating the parameter differences between pre- and post-unlearning models as gradients of unlearned images [18]. Du et al. [32] extend this idea to textual classification, enabling the reconstruction of unlearned text. Ma et al. [33] propose an RL-based approach that optimizes latent noise to reconstruct unlearned images. For LLMs, Wu et al. [19] extract unlearned text in LLMs by leveraging memorization and probability gaps between original and unlearned models, assuming the adversary knows the identifiable prefix of the target unlearned text. However, optimizing-based methods [18, 32, 33] rely on labels, which are unavailable in generative LLMs, and prefix-based methods rely on the strong assumption of access to sample-specific context. Currently, no unlearning inversion approaches are available without prior information about unlearned data.

3 Threat Model

Scenario. Following prior unlearning inversion attacks [18, 32, 33], we consider practical model update workflows where both pre- and post-unlearning model versions are accessible. In Machine Learning as a Service (MLaaS) settings [18], historical model versions are often retained for rollback, auditing, debugging, or staged deployment, introducing a side channel, as discrepancies induced by unlearning may encode information about removed training data. A broader discussion of practical deployment scenarios is provided in Appendix B.1.

Adversary. The adversary observes both pre- and post-unlearning models with white-box access to parameters, but has no access to the identities or content of the removed samples, the precise gradients or intermediate activations of the removed samples during training. The goal is to reconstruct samples removed through unlearning.

4 Method

4.1 TRACE Overview

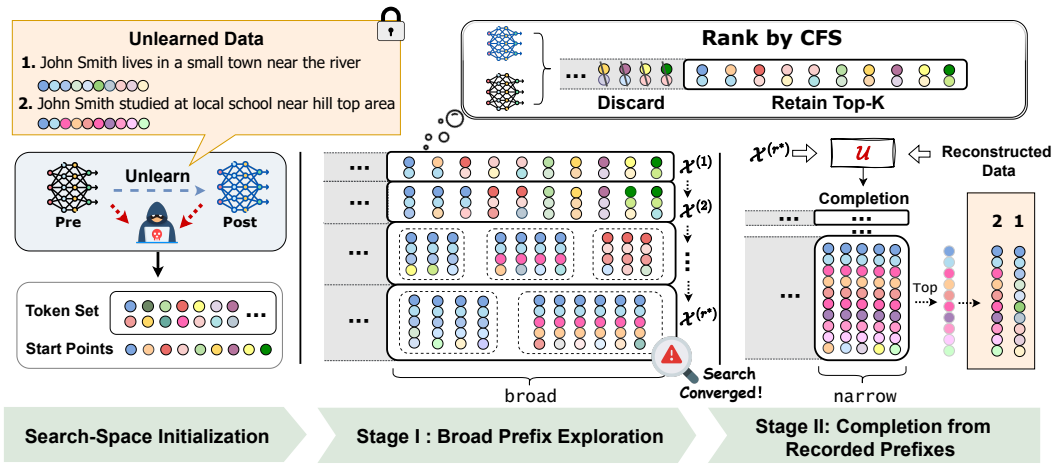


Figure 1: Framework of TRACE.

TRACE reconstructs the unlearned data by incrementally assembling sequences from a compact set of candidate tokens, which are derived from embedding differences between the pre- and post-unlearning models via the search-space initialization (Sec. 4.2). Reconstruction follows an iterative search process, where candidate sequences are iteratively extended, scored by a novel contrastive fluency score (CFS)(Sec 4.3), and pruned to retain the highest-scoring candidates. Building on the CFS-guided search, TRACE further structures the reconstruction process into two stages to facilitate the reconstruction of distinct unlearned samples (Sec 4.4): (1) **Stage I** - broad prefix exploration, which maintains a large set of candidate sequences at each step; (2) **Stage II** - completion from recorded prefixes, which completes the recorded prefixes from **Stage I** into full sequences with a relatively small candidate set at each step. The framework of TRACE is shown in Fig. 1.

4.2 Search-Space Initialization

Search-Space Restriction via Candidate Tokens. The large vocabulary of LLMs leads to a combinatorial explosion in possible sequences. Inspired by prior findings that non-zero embedding gradients reveal tokens in training data [27, 34], we identify an analogous effect in unlearning: tokens involved in the unlearning process exhibit non-zero embedding parameter differences between the pre-unlearning model θ^* and post-unlearning model θ_u. We exploit this property to construct a compact candidate token set to constrain the search space. Let $E \in \mathbb{R}^{|\mathcal{V}| \times h}$ denote the input embedding matrix, where $|\mathcal{V}|$ is the vocabulary size and h is the embedding dimension. For each token $v \in \mathcal{V}$, we compute the ℓ_2 norm of the embedding difference $\Delta(v)$ and define the candidate token set $\mathcal{V}_f$ by:

$$\mathcal{V}_f = \{v \in \mathcal{V} \mid \Delta(v) > 0\}, \quad \Delta(v) = \|E_{\theta^*}(v) - E_{\theta_u}(v)\|_2 . \tag{1}$$

Initialization of Starting Points. We construct a set of valid initial tokens $\mathcal{V}_0$ by filtering the candidate token set $\mathcal{V}_f$ based on linguistic constraints. Specifically, we retain those that are consistent with common sentence-initial patterns, such as tokens starting with uppercase letters or valid opening punctuation (e.g., ", (, [, {, or <).

4.3 Unified CFS-Guided Search

TRACE follows an iterative search process, where candidate sequences are iteratively extended, scored using a novel contrastive fluency score (CFS), and pruned to retain the highest-scoring candidates for next step. Let $\mathcal{X}^{(0)}$ denote the set of starting points, and $\mathcal{X}^{(r-1)}$ denote the retained set from step $r-1$. We construct the extended candidate set at step r by:

$$\mathcal{C}^{(r)} = \{\, x \oplus v \mid x \in \mathcal{X}^{(r-1)},\ v \in \mathcal{V}_f \,\}, \tag{2}$$

where $x \oplus v$ denotes the sequence obtained by extending x with token v. For an extended candidate sequence $x \in \mathcal{C}^{(r)}$, the perplexity [35] under a model θ is given by:

$$\mathrm{PPL}_\theta(x) = \exp\left(-\frac{1}{|x|}\sum_{i=1}^{|x|} \log p_\theta(x_i \mid x_{<i})\right). \tag{3}$$

Lower perplexity indicates that the sequence x has higher likelihood. Unlearning reduces the likelihood of target data in the post-unlearning model, resulting in a divergence in perplexity between the pre- and post-unlearning models. We leverage this to design a contrastive fluency score (CFS) to guide the search. Specifically, for a candidate sequence $x \in \mathcal{C}^{(r)}$, the contrastive score is defined as:

$$\mathrm{CFS}(x) = \frac{\mathrm{PPL}_{\theta_u}(x) - \mathrm{PPL}_{\theta^*}(x)}{\mathrm{PPL}_{\theta^*}(x)^{A(|x|)}}, \quad A(|x|) = \alpha + \beta \log |x|. \tag{4}$$

The length-aware exponent $A(|x|)$ controls the relative importance of the fluency constraint PPL_{θ^*}, where α and β are positive constants. This design is motivated by the fact that perplexity varies significantly with sequence length. As shown in Figure 2a, perplexity under the original model tends to be higher for shorter prefixes of unlearned data and decreases and stabilizes as the sequence grows. By progressively increasing the influence of $\mathrm{PPL}_{\theta^*}(x)$, CFS remains tolerant of high-perplexity prefixes during early exploration when sufficiently strong contrastive signals (i.e., a large perplexity gap ΔPPL) are present. This prevents premature pruning of promising sequences, while enforcing stricter fluency constraints in later steps. At each step, candidates in $\mathcal{C}^{(r)}$ are ranked according to $\mathrm{CFS}(x)$, and a subset is retained to form $\mathcal{X}^{(r)}$, which serves as the input for the next iteration. The specific retention strategy is detailed in Sec. 4.4.

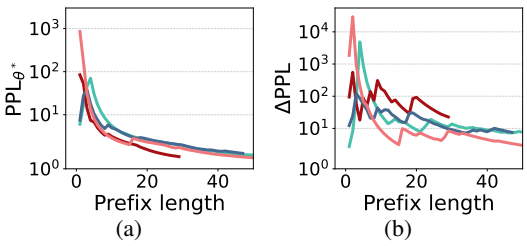


Figure 2: Effect of prefix length on perplexity for samples unlearned using GA. (a) Perplexity of unlearned samples under the pre-unlearning model. (b) Perplexity gap between pre- and post-unlearning models.

4.4 Two-Stage Reconstruction via CFS-Guided Search

While CFS defines a scoring principle for candidate sequences, reconstructing multiple distinct unlearned samples remains challenging, as sequential search tends to concentrate on a number of highly similar high-scoring sequences. To address this issue, TRACE proposes a two-stage architecture that separates diversity exploration from sequence completion. The procedure is summarized in Algorithm 1 in Appendix.

Stage I: Broad Prefix Exploration. In this stage, TRACE aims to explore diverse candidate prefixes that may correspond to different unlearned samples. To this end, TRACE maintains a relatively large retained set at each step. As the unlearning signal $\Delta\mathrm{PPL}$ exhibits strong sample-to-sample variability, particularly for short prefixes (Fig. 2b), TRACE ranks candidates in $\mathcal{C}^{(r)}$ under two parameterizations of CFS: $\mathrm{CFS}(\cdot;\alpha_s,\beta)$ and $\mathrm{CFS}(\cdot;\alpha_l,\beta)$, where α_s and α_l correspond to weaker and stronger fluency constraints, respectively, with $\alpha_s < \alpha_l$. The former is more permissive to high-perplexity prefixes in early steps and thus preserves candidates with strong unlearning signals, while the latter favors more fluent candidates, even if their unlearning signals are relatively weak. Accordingly, the retained set at step r is given by:

$$\mathcal{X}^{(r)} = \mathrm{Top}_k\Big(\mathcal{C}^{(r)};\mathrm{CFS}(\cdot;\alpha_s,\beta)\Big) \cup \mathrm{Top}_k\Big(\mathcal{C}^{(r)};\mathrm{CFS}(\cdot;\alpha_l,\beta)\Big). \tag{5}$$

To quantify diversity, sequences in $\mathcal{X}^{(r)}$ are grouped into clusters based on ROUGE-L similarity with a threshold τ. Formally, $\mathcal{X}^{(r)}$ is partitioned into clusters $\mathcal{P}^{(r)}$ such that

$$\forall P \in \mathcal{P}^{(r)},\ \forall x, x' \in P,\ \text{ROUGE-L}_{\text{recall}}(x, x') \geq \tau. \tag{6}$$

This stage continues until the retained set exhibits limited diversity, i.e., $|\mathcal{P}^{(r)}| \leq N$, where N serves as an approximate indication of how many distinct samples the attacker aims to reconstruct. The last retained set is recorded as $\mathcal{X}^{(r^*)}$, serving as candidate prefixes, which are then completed in Stage II. We use a broad search width $k = N \times d$ in stage I, where d is an expansion factor, to increase the chance of retaining prefixes corresponding to true unlearned samples.

Stage II: Completion from Recorded Prefixes. This stage revisits the last recorded prefix set $\mathcal{X}^{(r^*)}$ from Stage I and completes it into full sequences via a narrower CFS-guided search, where the strategy of extending and pruning candidates at each step follows Eq. (2) and Eq. (5) but with a smaller search width $k = Q$. To avoid redundant reconstruction, let $\mathcal{S}$ denote the set of reconstructed sequences. The set of uncovered prefixes $\mathcal{U}$ is:

$$\mathcal{U} = \left\{ x \in \mathcal{X}^{(r^*)} \;\middle|\; \max_{s\in\mathcal{S}} \text{ROUGE-L}_{\text{recall}}(x, s) < \tau \right\}. \tag{7}$$

TRACE performs iterative CFS guided search over $\mathcal{U}$. In each iteration, the search continues until a termination condition is met, yielding newly reconstructed sequences and triggering corresponding updates to $\mathcal{S}$ and $\mathcal{U}$. This process repeats until $\mathcal{U} = \emptyset$, i.e., all prefixes in $\mathcal{X}^{(r^*)}$ are covered. Detailed stopping criteria are provided in Appendix B.3.

5 Experiments

5.1 Setup

We evaluate TRACE on various standard unlearning benchmarks, including TOFU [36], MUSE [37], and WMDP [38]. For each dataset, we first fine-tune the model to obtain a pre-unlearning checkpoint, and then apply six representative unlearning methods, including gradient ascent (GA), gradient difference (GD), KL minimization (KL), negative preference optimization (NPO), NPO-Diff [36], and NPO-KL [36], to unlearn the target samples and obtain the corresponding unlearned models. Experiments are conducted on multiple open-source LLMs, including LLaMA2-13B, LLaMA2-7B [39],Mistral-7B [40], Qwen2.5-3B [41], StableLM-3B-4E1T [42] and TinyLlama-1.1B [43]. Follow [19], three metrics are adopted for evaluation, including ROUGE-L recall (R), and average extraction success rate (A-ESR$_\kappa$), which measures the percentage of ground-truth unlearned samples that can be matched by reconstructed samples under the ROUGE-L recall threshold $\kappa \in \{1.0, 0.9\}$ (as formally defined in Appendix D.2). We set $N = 1.5 \times B$ in our experiments, where B is the unlearning batch size. TRACE also involves the CFS coefficients $(\alpha_s, \alpha_l, \beta)$, similarity threshold τ, and search parameters d and Q. We analyze their sensitivity in Sec. 5.4; complete experimental settings and detailed explanations of their roles are provided in Appendices D.3 and B.2, respectively.

Table 1: Reconstruction performance on TOFU and MUSE.

Dataset	Model	Metric	GA		GD		KL		NPO		NPO-Diff		NPO-KL	
			Ours	PPL-only	Ours	PPL-only	Ours	PPL-only	Ours	PPL-only	Ours	PPL-only	Ours	PPL-only
TOFU	TinyLlama-1.1B	ROUGE-L(R)	0.9084	0.4383	0.8770	0.1770	0.8785	0.1770	0.8882	0.4383	0.8358	0.1770	0.8472	0.1770
		A-ESR$_{1.0}$(%)	78.75	10.00	71.25	2.50	76.25	2.50	76.25	10.00	68.75	2.50	67.50	2.50
		A-ESR$_{0.9}$(%)	86.25	15.00	82.50	2.50	82.50	2.50	82.50	15.00	76.25	2.50	75.00	2.50
	Qwen2.5-3B	ROUGE-L(R)	0.8409	0.1681	0.8191	0.1387	0.8362	0.1386	0.8189	0.1592	0.8259	0.1534	0.8625	0.1586
		A-ESR$_{1.0}$(%)	67.50	0.00	68.75	0.00	66.25	0.00	65.00	0.00	67.50	0.00	71.25	0.00
		A-ESR$_{0.9}$(%)	76.25	1.25	75.00	1.25	76.25	1.25	75.00	1.25	76.25	1.25	78.75	1.25
	LLaMA2-7B	ROUGE-L(R)	0.9799	0.6017	0.9695	0.2946	0.9592	0.2946	0.9715	0.6017	0.9578	0.2946	0.9684	0.2946
		A-ESR$_{1.0}$(%)	95.00	28.75	92.50	7.50	90.00	7.50	93.75	28.75	93.75	7.50	96.25	7.50
		A-ESR$_{0.9}$(%)	97.50	36.25	96.25	10.00	91.25	10.00	95.00	36.25	95.00	10.00	96.25	10.00
	Mistral-7B	ROUGE-L(R)	0.9594	0.7145	0.9873	0.5584	0.8707	0.5584	0.9602	0.7145	0.9604	0.5584	0.9299	0.5584
		A-ESR$_{1.0}$(%)	90.00	47.50	92.50	37.50	61.25	37.50	90.00	47.50	88.75	37.50	82.50	37.50
		A-ESR$_{0.9}$(%)	95.00	51.25	98.75	37.50	67.50	37.50	95.00	51.25	93.75	37.50	86.25	37.50
	LLaMA2-13B	ROUGE-L(R)	0.9936	0.6106	0.9792	0.3311	0.9792	0.3311	0.9795	0.6106	0.9703	0.3311	0.9690	0.3311
		A-ESR$_{1.0}$(%)	93.75	33.75	92.50	10.00	93.75	10.00	96.25	33.75	93.75	10.00	93.75	10.00
		A-ESR$_{0.9}$(%)	98.75	40.00	97.50	10.00	97.50	10.00	96.25	40.00	95.00	10.00	95.00	10.00
MUSE	TinyLlama-1.1B	ROUGE-L(R)	0.8750	0.2834	0.7555	0.1153	0.7641	0.1153	0.8467	0.2834	0.8136	0.1153	0.8099	0.1153
		A-ESR$_{1.0}$(%)	80.00	17.50	65.00	3.75	67.50	3.75	76.25	17.50	71.25	3.75	76.25	3.75
		A-ESR$_{0.9}$(%)	83.75	20.00	67.50	5.00	70.00	5.00	78.75	20.00	75.00	5.00	77.50	5.00
	Stablelm-3B	ROUGE-L(R)	0.8633	0.5624	0.7448	0.4246	0.7883	0.4246	0.8838	0.5624	0.7958	0.4246	0.7728	0.4246
		A-ESR$_{1.0}$(%)	80.00	46.25	62.50	32.50	72.50	32.50	83.75	46.25	70.00	32.50	68.75	32.50
		A-ESR$_{0.9}$(%)	85.00	51.25	71.25	36.25	78.75	36.25	87.50	51.25	76.25	36.25	75.00	36.25
	LLaMA2-7B	ROUGE-L(R)	0.9457	0.3915	0.9332	0.1654	0.9398	0.1654	0.9579	0.3915	0.9554	0.1654	0.9646	0.1654
		A-ESR$_{1.0}$(%)	93.75	37.50	92.50	13.75	93.75	13.75	95.00	37.50	93.75	13.75	96.25	13.75
		A-ESR$_{0.9}$(%)	93.75	37.50	92.50	13.75	93.75	13.75	95.00	37.50	93.75	13.75	96.25	13.75
	Mistral-7B	ROUGE-L(R)	0.9549	0.3533	0.8753	0.1221	0.9760	0.1221	0.9637	0.3533	0.9540	0.1221	0.9604	0.1221
		A-ESR$_{1.0}$(%)	93.75	31.25	83.75	6.25	96.25	6.25	94.44	31.25	90.00	6.25	93.75	6.25
		A-ESR$_{0.9}$(%)	93.75	32.50	85.00	7.50	96.25	7.50	94.44	32.50	93.75	7.50	93.75	7.50
	LLaMA2-13B	ROUGE-L(R)	0.9327	0.3839	0.9438	0.1728	0.9628	0.1728	0.9441	0.3839	0.9399	0.1728	0.9513	0.1728
		A-ESR$_{1.0}$(%)	92.50	35.00	92.50	12.50	96.25	12.50	92.50	35.00	92.50	12.50	95.00	12.50
		A-ESR$_{0.9}$(%)	92.50	36.25	92.50	13.75	96.25	13.75	92.50	36.25	92.50	13.75	95.00	13.75

5.2 Multi-sample reconstruction without prior knowledge of unlearned data

We evaluate TRACE against a PPL-only baseline that adopts the same framework as TRACE but ranks candidates solely by perplexity under the pre-unlearning model. This baseline serves as an ablation to isolate the contribution of the CFS mechanism from the original model's inherent memorization. In Appendix D.4, we further compare PPL-only with FILM [27], and show that PPL-only is a stronger reconstruction baseline. Table 1 summarizes the results averaged over 10 randomly sampled batches with an unlearning size of $B = 8$. TRACE consistently achieves superior performance across various models and unlearning methods. While the efficacy of PPL-only deteriorates significantly under unlearning methods that incorporate retain-set regularization (e.g., GD, KL, NPO-Diff/KL), TRACE remains remarkably robust across all such settings. Furthermore, we observe that larger models are more susceptible to reconstruction than smaller ones due to their superior capacity for memorization. For instance, TRACE achieves ROUGE-L(R) above 0.93 on LLaMA2-7B, Mistral-7B, and LLaMA2-13B. We provide a visualization of reconstructed data in Appendix E.1.

Table 2 further evaluates different unlearning batch sizes $B \in \{8, 16, 32, 128\}$. TRACE remains effective even at $B = 128$, achieving ROUGE-L(R) scores above 0.8 on MUSE. Although performance decreases as batch size increases, this degradation can be mitigated by increasing expansion factor d, which allows more candidate sequences to be retained per step and improves the chance of capturing true prefixes corresponding to unlearned samples (see Section 5.4).

Table 2: TRACE Performance under different unlearning batch sizes using LLaMA2-7B.

Dataset	Metric	GA				NPO			
		$B = 8$	$B = 16$	$B = 32$	$B = 128$	$B = 8$	$B = 16$	$B = 32$	$B = 128$
TOFU	ROUGE-L(R)	0.9799	0.9827	0.9541	0.7924	0.9715	0.9569	0.9287	0.8146
	A-ESR$_{1.0}$(%)	95.00	92.50	88.75	71.09	93.75	87.50	87.50	73.44
	A-ESR$_{0.9}$(%)	97.50	97.50	92.50	74.22	95.00	92.50	90.62	75.78
MUSE	ROUGE-L(R)	0.9457	0.9408	0.8996	0.8007	0.9579	0.9561	0.8897	0.8051
	A-ESR$_{1.0}$(%)	93.75	92.50	87.81	74.22	95.00	93.75	86.56	74.22
	A-ESR$_{0.9}$(%)	93.75	92.50	88.12	77.34	95.00	94.38	87.18	81.25

5.3 Comparison with Prefix-Assisted Baseline UBNF [19]

To enable a fair comparison with the prefix-conditioned unlearned text reconstruction setting in UBNF [19], we conduct reconstruction on the long-form WMDP dataset, where the goal is to

reconstruct a long sensitive text. Each WMDP sample is a 512-token text, and UBNF [19] is given the first 10 words as a prefix, while TRACE uses none. As shown in Fig. 3, TRACE consistently outperforms both baselines across models and six unlearning methods without the strong assumptions in the known prefix. In particular, on LLaMA2-7B, TRACE improves ROUGE-L(R) by more than 100% over UBNF. These results demonstrate the effectiveness of TRACE in realistic long-form reconstruction scenarios. We provide a visualization of reconstructed text in Appendix E.2.

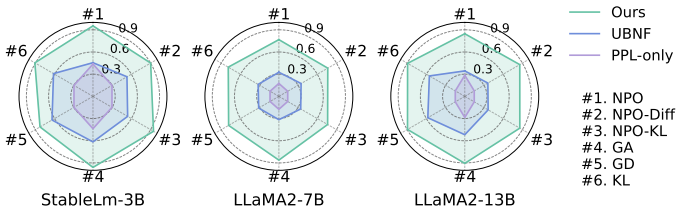


Figure 3: Long-form reconstruction performance on WMDP measured by ROUGE-L(R).

Study Case: Prefix-Matched Reconstruction on RWKU Dataset [44]. We further evaluate prefix-assisted reconstruction on the RWKU dataset, where unlearned samples have an average length of 150 words. We compare three methods given equal-length prefixes (2–10 words) from the unlearned data, using GA with an unlearning batch size of 16. As shown in Fig. 4, TRACE consistently outperforms UBNF, demonstrating its effectiveness under limited prefix information in recovering the remaining content.

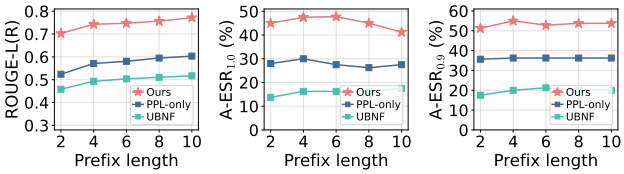


Figure 4: Reconstruction performance with prefix on RWKU using LLaMA2-7B

5.4 Ablation Study

Hyperparameters α_s, α_l and β in CFS. Figure 5 shows the sensitivity of CFS to α_s, α_l and β on MUSE using TinyLlama under NPO. Introducing the length-aware term $\beta > 0$ improves the robustness of CFS, preventing performance collapse at small α_s, with the best results around $\alpha_s = 2$. This suggests a beneficial dynamic trade-off: early steps prioritize strong unlearning signals even if prefixes are not fluent, while later steps increasingly favor linguistic plausibility as sequences become more complete. Regarding α_l, setting $\alpha_l = \alpha_s = 2$ degrades performance as the dual-ranking reverts to a single objective. $\alpha_l \in [3, 8]$ leads to stronger performance, suggesting that this branch serves as a complement to the small α_s setting by capturing fluent prefixes with weaker unlearning signals that may otherwise be overlooked.

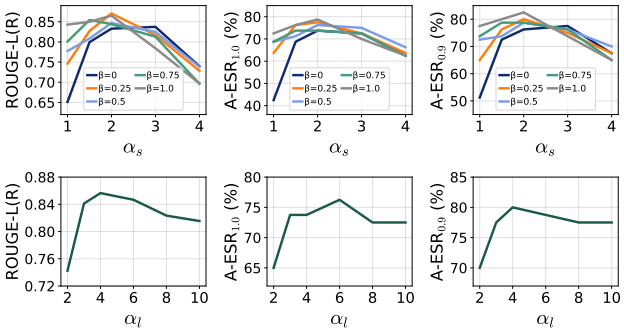


Figure 5: Hyperparameters in CFS on MUSE. Top: varying α_s and β ($\alpha_l = 6.0$). Bottom: varying α_l ($\alpha_s = 2, \beta = 0.5$).

Parameter d in stage I. The parameter d determines how many candidates are retained at each step in stage I and therefore directly affects the ability to capture diverse prefixes corresponding to different unlearned samples. As observed in Table 2, increasing the unlearning batch size B to 32 (with a fixed $d = 5$) leads to performance degradation on the MUSE dataset. Table 3 reports that increasing d improves reconstruc-

Table 3: Effect of d on MUSE.

Metric	$d = 5$	$d = 10$	$d = 15$
ROUGE-L(R)	0.8897	0.9504	0.9564
A-ESR$_{1.0}$(%)	86.56	92.50	92.81
A-ESR$_{0.9}$(%)	87.19	93.13	93.75

tion performance across metrics, indicating that TRACE can scale to larger batch unlearning settings by increasing d.

Parameter Q in stage II. The parameter Q controls the search width in Stage II. As the search focuses on extending selected prefixes rather than exploring diverse trajectories, a narrow search width is sufficient. As shown in Table 4, a relatively small value (e.g., $Q = 5$) achieves strong performance. These results highlight the effectiveness of decoupling diversity discovery from sequence completion, enabling Stage II to achieve reliable reconstruction with a modest search budget while avoiding redundant exploration.

Table 4: Effect of Q on MUSE.

Metric	$Q = 1$	$Q = 5$	$Q = 10$
ROUGE-L(R)	0.8148	0.8529	0.8467
A-ESR$_{1.0}$(%)	68.75	76.25	76.25
A-ESR$_{0.9}$(%)	76.25	78.75	77.50

Effect of Similarity Threshold τ. We study the impact of the similarity threshold τ used in Eq. (6) and Eq. (7), using LLaMA2-7B under NPO with unlearning batch size $B = 128$. As shown in Fig. 6, Reconstruction performance on TOFU demonstrates a higher sensitivity to the choice of τ compared to that on MUSE. We attribute this to its structural properties: TOFU consists of question–answer pairs, where samples tend to share higher lexical and semantic similarity. When τ is too small, structurally similar but semantically different prefixes are merged into the same cluster, leading to inadequate reconstruction of distinct samples. Empirically, setting $\tau \in [0.6, 0.7]$ is found to provide a robust balance between distinguishing distinct sequences and merging redundant sequences.

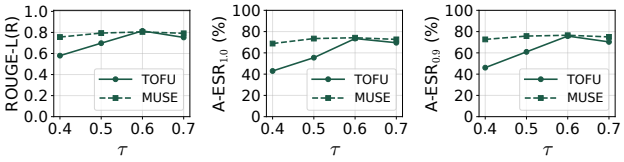


Figure 6: Effect of Similarity Threshold τ

Training and Unlearning Epochs. Figure 7 presents the impact of training and unlearning epochs on reconstruction performance on MUSE using TinyLlama under NPO. We observe that varying the number of unlearning epochs has a minimal effect on reconstruction performance. In contrast, increasing the number of training epochs consistently enhances reconstruction quality, suggesting that stronger pre-unlearning memorization makes the target data more recoverable. Notably, our method TRACE consistently outperforms the baseline across all settings.

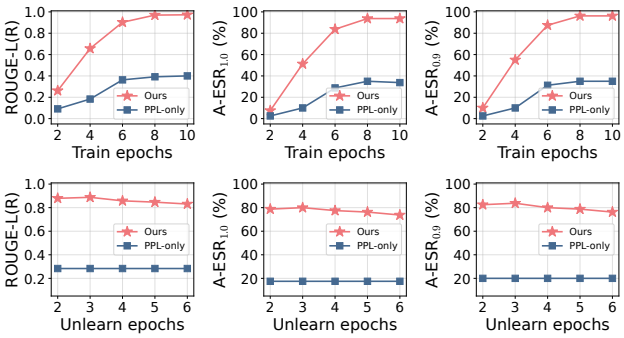


Figure 7: Effect of training and unlearning epochs. Top: varying training epochs (unlearning epochs = 5). Bottom: varying unlearning epochs (training epochs = 5).

5.5 TRACE in the Black-Box Setting

Our main experiments consider a white-box setting, where TRACE exploits embedding differences between the pre- and post-unlearning models to construct a compact candidate token set V_f. We further investigate whether TRACE can operate under a weaker black-box setting, where model parameters are inaccessible but token-level output probabilities from both model versions are available. In this setting, we therefore set V_f to the entire vocabulary and perform reconstruction. We evaluate this setting on MUSE using TinyLlama under NPO with an unlearning batch size of 8, considering pre-unlearning models trained for 4, 5, and 6 epochs.

As shown in Table 5, TRACE retains meaningful reconstruction capability in the black-box setting, although its performance is lower than in the white-box setting. Without parameter access, TRACE must instead search over the entire vocabulary, introducing many competing candidate prefixes and making it more difficult to retain the ground-truth paths during exploration. This challenge would become increasingly pronounced for models with larger vocabularies. In such a large search space,

Table 5: TRACE performance in the black-box setting.

Dataset	Metric	Training epochs = 4		Training epochs = 5		Training epochs = 6	
		white-box	black-box	white-box	black-box	white-box	black-box
MUSE	ROUGE-L(R)	0.6566	0.3419	0.8467	0.6043	0.9027	0.6274
	A-ESR$_{1.0}$ (%)	51.25	27.50	76.25	55.00	83.75	57.50
	A-ESR$_{0.9}$ (%)	55.00	27.50	78.75	55.00	87.50	60.00

stronger memorization of the target samples exposes them to greater reconstruction risk, as evidenced by the substantial improvement in reconstruction performance with increasing pre-unlearning training epochs.

5.6 Robustness to Checkpoint Transformations

In practical deployments, model checkpoints may undergo additional transformations after unlearning. We evaluate TRACE under two representative model compression techniques, including low-bit quantization and weight pruning.

Model Quantization. We apply 4-bit AWQ quantization [45] to the unlearned LLaMA2-7B model obtained with NPO on MUSE, and evaluate TRACE using the resulting quantized checkpoint. As shown in Table 6, quantization has only a minor effect on reconstruction performance: ROUGE-L(R) decreases from 0.9579 to 0.9506, while both A-ESR$_{1.0}$ and A-ESR$_{0.9}$ remain at 95%. This result shows that the reconstruction signal exploited by TRACE remains largely preserved after low-bit quantization.

Table 6: TRACE under 4-bit AWQ quantization.

Metric	Original	AWQ 4-bit
ROUGE-L(R)	0.9579	0.9506
A-ESR$_{1.0}$ (%)	95.00	95.00
A-ESR$_{0.9}$ (%)	95.00	95.00

Model Pruning. We apply Wanda [46] with different pruning ratios to the unlearned LLaMA2-7B model obtained with NPO on MUSE, and evaluate TRACE using the resulting pruned checkpoints. Table 7 shows that TRACE remains effective under moderate pruning, with ROUGE-L(R) decreasing only slightly from 0.9579 to 0.9503 at a 10% pruning ratio. More aggressive pruning progressively reduces reconstruction quality, with ROUGE-L(R) decreasing to 0.8489 and 0.6394 at pruning ratios of 30% and 50%, respectively. Interestingly, aggressive pruning also substantially decreases the perplexity of the unlearned samples, from 16.07 at 0% pruning to 5.07 at 50% pruning. Thus, while aggressive pruning weakens TRACE's reconstruction performance, it simultaneously compromises the effect of unlearning.

Table 7: TRACE under Wanda pruning.

Pruning Ratio	PPL	ROUGE-L(R)
0%	16.07	0.9579
10%	17.50	0.9503
30%	13.58	0.8489
50%	5.07	0.6394

6 Possible Defenses

We evaluate three practical defenses: *parameter obfuscation* [47], *gradient pruning* [48], and *post-unlearning fine-tuning* [18].

Parameter Obfuscation. Inspired by DP-SGD [47], we adopt a lightweight parameter obfuscation mechanism during unlearning. At each unlearning step, we first clip gradients to $C = 1.2$, and then inject noise proportional to the gradient norm, i.e., $\tilde{g} = g + \sigma\|g\|\frac{z}{\|z\|}, \quad z \sim \mathcal{N}(0, I)$, with hyperparameter σ controlling the perturbation strength. We emphasize that this approach is heuristic and does not aim to provide formal differential privacy guarantees. The injected noise perturbs the embedding parameters across the entire vocabulary, which interferes with our candidate token set V_f extraction. To address this, we apply a simple unsupervised refinement. Specifically, we first select the top-$p\%$ tokens ($p = 20$ in our setting) with the largest embedding differences, and then perform 1D clustering (K-means with $k = 2$) on their log-scaled differences, taking the cluster with larger mean shifts as V_f. Table 8a reports the trade-off between unlearning effectiveness measured by perplexity (higher is better) and reconstruction performance under different noise levels σ on MUSE using TinyLlama under GA. As σ increases, perplexity on the forget set decreases, indicating weaker

Table 8: Possible defenses. Rec($\mathcal{V}_f$) and Prec($\mathcal{V}_f$) denote recall and precision over V_f with respect to ground-truth tokens from the unlearned data.

(a) Parameter Obfuscation.

σ	Perplexity	ROUGE-L(R)	Rec($\mathcal{V}_f$)	Prec($\mathcal{V}_f$)
0	20.17	0.8750	1	1
0.01	19.37	0.8361	0.9551	1
0.05	18.63	0.7226	0.9038	1
0.10	17.81	0.5375	0.8525	1
0.15	17.06	0.4599	0.8846	0.0601
0.20	16.30	0.4062	0.8525	0.0562

(b) Gradient Pruning.

p	Perplexity	ROUGE-L(R)	Rec($\mathcal{V}_f$)	Prec($\mathcal{V}_f$)
0	20.17	0.8750	1	1
0.80	25.73	0.8750	1	1
0.85	19.43	0.8750	1	1
0.90	12.81	0.8636	1	1
0.95	5.57	0.8636	1	1
0.99	2.29	0.8182	1	1

unlearning. Meanwhile, the precision of V_f sharply deteriorates when $\sigma \geq 0.15$, so that a large number of noise-induced tokens become indistinguishable from ground-truth tokens. For smaller noise levels ($\sigma \leq 0.10$), TRACE remains effective with high precision and strong recall of V_f.

Gradient Pruning. We evaluate gradient pruning [48] on MUSE using TinyLlama under GA with pruning ratios $p \in \{0.8, 0.85, 0.9, 0.95, 0.99\}$, where a fraction p of gradient components with the smallest magnitudes are pruned (set to zero) during each update. Pruning has little impact on TRACE, with the extraction of V_f remaining unaffected. As shown in Table 8b, even with an aggressive pruning ratio of 0.99, TRACE still achieves an average ROUGE-L (R) of 0.8182.

Fine-tuning. Following the prior work [18], we consider a setting where the developer fine-tunes the unlearned model before uploading the updated model. Specifically, we fine-tune the unlearned model (obtained via GA) for two epochs on retain sets of varying sizes 0, 100, 200, 300, 400, and 500 on the MUSE dataset using LLaMA2-7B. As shown in Fig. 8, fine-tuning partially reduces reconstruction performance as the size of the fine-tuning dataset increases. This process enlarges the token candidate set V_f and expands the search space, making reconstruction more challenging. Nevertheless, TRACE consistently outperforms the PPL-only baseline.

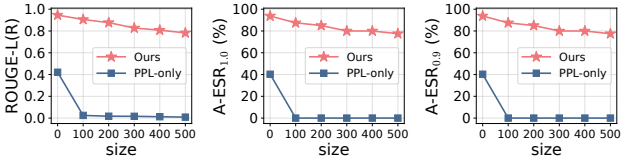


Figure 8: Fine-tuning after unlearning.

Discussion. These results suggest that heuristic defenses provide limited protection against TRACE. Gradient pruning has little effect on reconstruction. Parameter obfuscation can reduce attack effectiveness, but at the cost of weaker unlearning performance; Post-unlearning fine-tuning partially degrades reconstruction by expanding the search space. However, it does not eliminate the perplexity gap induced by unlearning, and TRACE can still be applied, albeit at the cost of increased computational overhead (analyzed in Appendix C), which is often acceptable in offline attack settings. We further investigate whether exact retraining can provide stronger resistance to TRACE in Appendix D.5. The results show that exact retraining substantially reduces the attack effectiveness compared with approximate unlearning, although a small fraction of the unlearned samples can still be exactly reconstructed by TRACE. Overall, these findings underscore the need for unlearning methods and deployment protocols that explicitly account for the attack surface introduced by approximate unlearning.

7 Conclusion

In this work, we reveal that approximate unlearning in LLMs leaves recoverable footprints in model snapshots, which are commonly retained for auditing or rollback, facilitating the privacy breaches it is intended to prevent even without requiring any data-specific prior knowledge. To demonstrate this, we propose TRACE, a framework that reconstructs unlearned text by utilizing parameter differences to effectively constrain the vast search space and employing a contrastive search strategy within a two-stage architecture. Extensive experiments show that TRACE consistently achieves high-fidelity reconstruction across diverse models, datasets, and unlearning methods. This highlights the need for unlearning methods and deployment protocols that explicitly account for parameter-difference leakage, and call for future work on defenses that mitigate cross-snapshot information exposure.

References

[1] Gdpr article 17: Right to erasure. gdpr-info.eu, 2016. URL `https://gdpr-info.eu/art-17-gdpr/`.

[2] Regulation (eu) 2024/1689: Artificial intelligence act. Official Journal of the European Union, 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`.

[3] Denghui Zhang, Zhaozhuo Xu, and Weijie Zhao. Llms and copyright risks: Benchmarks and mitigation approaches. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 44–50, 2025.

[4] Kalin Hristov. Artificial intelligence and the copyright dilemma. *Idea*, 57:431, 2016.

[5] Felix B Mueller, Rebekka Görge, Anna K Bernzen, Janna C Pirk, and Maximilian Poretschkin. Llms and memorization: On quality and specificity of copyright compliance. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 984–996, 2024.

[6] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On protecting the data privacy of large language models (llms) and llm agents: A literature review. *High-Confidence Computing*, 5(2):100300, 2025.

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

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

[9] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

[10] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, 2025.

[11] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 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, 2023.

[12] XiaoHua Feng, Chaochao Chen, Yuyuan Li, and Zibin Lin. Fine-grained pluggable gradient ascent for knowledge unlearning in language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 10141–10155, 2024.

[13] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[14] Zirui Pang, Hao Zheng, Zhijie Deng, Ling Li, Zixin Zhong, and Jiaheng Wei. Label smoothing improves gradient ascent in llm unlearning. *arXiv preprint arXiv:2510.22376*, 2025.

[15] Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR, 2022.

[16] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

[17] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

[18] Hongsheng Hu, Shuo Wang, Tian Dong, and Minhui Xue. Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3257–3275. IEEE, 2024.

[19] Xiaoyu Wu, Yifei Pang, Terrance Liu, and Zhiwei Steven Wu. Unlearned but not forgotten: Data extraction after exact unlearning in llm. *arXiv preprint arXiv:2505.24379*, 2025.

[20] Jie Ren, Yue Xing, Yingqian Cui, Charu C Aggarwal, and Hui Liu. Sok: Machine unlearning for large language models. *arXiv preprint arXiv:2506.09227*, 2025.

[21] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1817–1829, 2024.

[22] Yoichi Ishibashi and Hidetoshi Shimodaira. Knowledge sanitization of large language models. *arXiv preprint arXiv:2309.11852*, 2023.

[23] Tianyang Xu, Xiaoze Liu, Feijie Wu, Xiaoqian Wang, and Jing Gao. Suv: Scalable large language model copyright compliance with regularized selective unlearning. *arXiv preprint arXiv:2503.22948*, 2025.

[24] Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana R Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *Advances in Neural Information Processing Systems*, 37:12581–12611, 2024.

[25] James Y Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models. *arXiv preprint arXiv:2404.11045*, 2024.

[26] Mislav Balunovic, Dimitar Dimitrov, Nikola Jovanović, and Martin Vechev. Lamp: Extracting text from gradients with language model priors. *Advances in Neural Information Processing Systems*, 35:7641–7654, 2022.

[27] Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. Recovering private text in federated learning of language models. *Advances in neural information processing systems*, 35:8130–8143, 2022.

[28] Ivo Petrov, Dimitar I Dimitrov, Maximilian Baader, Mark N Müller, and Martin Vechev. Dager: Exact gradient inversion for large language models. *Advances in neural information processing systems*, 37:87801–87830, 2024.

[29] Jin Xie, Ruishi He, Songze Li, Xiaojun Jia, and Shouling Ji. Recit: Reconstructing full private data from gradient in parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2504.20570*, 2025.

[30] Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390, 2020.

[31] Chenxi Dai, Lin Lu, and Pan Zhou. Stealing training data from large language models in decentralized training through activation inversion attack. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14539–14551, 2025.

[32] Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. Textual unlearning gives a false sense of unlearning. *arXiv preprint arXiv:2406.13348*, 2024.

[33] Mengyao Ma, Shuofeng Liu, Jason Xue, Surya Nepal, and Guangdong Bai. Retrace: Reinforcement learning-guided reconstruction attacks on machine unlearning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[34] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[35] Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, (2):179–190, 1983.

[36] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

[37] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

[38] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[40] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

[41] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[42] Jonathan Tow, Marco Bellagente, Dakota Mahan, and Carlos Riquelme. Stablelm 3b 4e1t. *URL https://huggingface. co/stabilityai/stablelm-3b-4e1t*, 2023.

[43] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[44] Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, Jun Zhao, et al. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Advances in Neural Information Processing Systems*, 37:98213–98263, 2024.

[45] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

[46] Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964, 2024.

[47] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[48] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

[49] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. Pmlr, 2017.

[50] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Asynchronous federated optimization. *arXiv preprint arXiv:1903.03934*, 2019.

A Societal Impact and Potential Harmful Consequences

This work studies a privacy and security vulnerability in approximate machine unlearning for large language models. Our method, TRACE, demonstrates that when an adversary has access to both pre- and post-unlearning model snapshots, it may be possible to reconstruct data that was intended to be removed, even without any data-specific prior knowledge.

Potential positive societal impacts. This work aims to improve the safety of machine unlearning systems by revealing a previously underappreciated failure mode. We emphasize that it is intended for defensive and scientific use only, and to motivate more robust unlearning methods, improved snapshot management, and stronger privacy auditing practices. In practice, we recommend minimizing snapshot exposure, enforcing strict access control over model checkpoints, and combining unlearning with additional privacy-preserving techniques prior to deployment.

Potential negative societal impacts. The primary risk of this research is dual use. An adversary could potentially use the reconstruction procedure to recover sensitive training samples, including personally identifiable information, confidential documents, copyrighted text, or other data that a model owner intended to erase. Such misuse could undermine user privacy, weaken compliance with data deletion requests, and reduce trust in machine unlearning as a privacy-preserving mechanism.

Security considerations. We do not propose a mechanism for weaponization or harassment, nor do we study surveillance, discrimination, or any application that targets protected attributes. Our experiments are conducted on standard research benchmarks, and the paper does not involve human subjects or any direct collection of private user data.

B Details of TRACE

B.1 Practical Deployment Scenarios

Our threat model assumes access to both pre- and post-unlearning model versions, following prior unlearning inversion attacks [18, 32, 33]. While such access can naturally arise in practical model update workflows where multiple model versions are retained or distributed. We discuss representative deployment scenarios below to further motivate this access assumption.

Model Update and Maintenance. In MLaaS[18] and other continuously maintained model services, model updates may produce multiple versions that are retained for operational purposes, such as rollback, auditing, debugging, and staged deployment. Consequently, personnel involved in model maintenance, such as ML engineers, platform operators, or auditors, may have legitimate access to both the model before an unlearning request and its subsequently updated version. Such access enables comparison between the two model versions, potentially exposing information encoded in the parameter changes induced by unlearning.

Federated Learning. Federated learning [49], particularly asynchronous federated learning [50], provides a natural setting for dual-model access, where successive global model versions may reflect individual client unlearning updates. Suppose a global model M_t is available to participating clients, and the subsequent version M_{t+1} incorporates an unlearning update from a particular client. If no other client updates are incorporated between these two versions, the difference $M_{t+1} - M_t$ may predominantly reflect the effect of that unlearning operation. Consequently, the potential adversary is not limited to the aggregation server: any participant with access to both versions could potentially reconstruct the data removed by another client, without accessing that client's private local data.

API-Based Deployment. While TRACE primarily targets the white-box setting, we further explore its applicability in a black-box scenario where model providers expose different model versions through APIs to their customers while keeping their internal parameters private. Users with legitimate API access can query both versions and observe their outputs. When token-level output probabilities are available, these outputs provide the contrastive signal required by TRACE. We evaluate the effectiveness of TRACE in this setting in Sec. 5.5. Despite the increased reconstruction difficulty and

computational cost compared with the white-box setting, TRACE can still uncover some unlearned samples.

B.2 Detailed Explanation of Hyperparameters in TRACE

TRACE involves several hyperparameters that control different aspects of the reconstruction process. We clarify their roles and selection below. Importantly, these hyperparameters do not require prior knowledge of the unlearned data, including the unlearning batch size.

Reconstruction budget N. N serves as an approximate indication of how many distinct samples the attacker aims to reconstruct. It controls the target level of prefix diversity during Stage I. Specifically, TRACE groups the retained candidate prefixes according to their similarity, and N determines how many distinct prefix groups the exploration stage aims to cover. A larger N can subsequently produce more diverse reconstructed sequences and increase reconstruction coverage. N can be specified independently, we set $N = 1.5B$ for controlled evaluation in our experiments. We evaluate TRACE with different values of N on MUSE using LLaMA2-7B under NPO. As shown in Table 9, increasing N generally improves reconstruction coverage.

Table 9: Effect of N on MUSE using LLaMA2-7B under NPO ($B = 8$).

N	8	10	12	14
ROUGE-L(R)	0.9407	0.9459	0.9579	0.9650
A-ESR$_{1.0}$ (%)	92.50	93.75	95.00	95.00
A-ESR$_{0.9}$ (%)	95.00	95.00	95.00	96.25

Search parameters d and Q. The parameters d and Q control the search width in Stage I and Stage II, respectively. In Stage I, TRACE uses $k = N \times d$, where d is an expansion factor that determines how broadly TRACE explores candidate prefixes. A larger d retains more prefixes and can improve reconstruction coverage, at the cost of additional computation. In Stage II, Q determines the search width for completing the selected prefixes. Since this stage focuses on sequence completion rather than diversity exploration, a relatively small Q is generally sufficient.

CFS parameters α_s, α_l, and β. The parameters α_s, α_l, and β determine how CFS balances the memorization signal of the original model and the memorization difference induced by unlearning. The ranges considered for these parameters follow their algorithmic roles rather than any knowledge of the unlearned samples. We summarize their roles and the ranges examined in our experiments in Table 10.

Table 10: Roles and evaluated ranges of the CFS hyperparameters.

Parameter	Test ranges	Rationale
α_s	1–4	$\alpha_s = 1$ is the natural starting point, where the CFS score reduces to $CFS(x) = \frac{PPL_{\theta_u}(x) - PPL_{\theta^*}(x)}{PPL_{\theta^*}(x)}$ at the first step.
α_l	2–10	$\alpha_l > \alpha_s$ compensates for the unstable unlearning signals of short prefixes by favoring fluent candidate sequences.
β	0–1	$\beta = 0$ ignores sequence length, while larger β progressively increases the influence of sequence length in CFS.

Similarity threshold τ. The threshold τ determines when candidate sequences are considered similar during prefix clustering and redundancy removal. A larger τ allows TRACE to distinguish and retain more mutually similar sequences. As shown in Figure 6, using a default value of $\tau = 0.6$ can achieve strong performance, regardless of whether the unlearned samples are highly similar (e.g., TOFU) or largely diverse (e.g., MUSE).

B.3 Detailed stopping criteria.

In Stage II, TRACE maintains the same dual ranking mechanism introduced in Stage I, based on $\mathrm{CFS}(\cdot; \alpha_s, \beta)$ and $\mathrm{CFS}(\cdot; \alpha_l, \beta)$. At each step, the top-ranked sequence under each scoring function

is tracked independently until it satisfies one of the stopping conditions defined below, after which it is added to the reconstructed set $\mathcal{S}$. The search terminates once both scoring functions have triggered a stopping condition, which may occur at different steps. Specifically, the top-ranked sequence is considered complete if any of the following holds:

End-of-sequence token. The last token is an end-of-sequence token (e.g., EOS).

Consecutive PPL Increases. When a sentence ending punctuation mark (e.g., .) is encountered, search terminates if PPL increases consecutively with subsequent sequence extensions.

Maximum length constraint. As a fallback, search terminates when the sequence length exceeds a predefined maximum length $L_{\max}$.

C Computational Overhead

Compared to standard autoregressive generation, TRACE incurs a higher computational cost. The computational cost per step in TRACE primarily consists of model forward passes and candidate ranking. At each step, while standard autoregressive generation involves a single forward pass, TRACE increases the number of forward passes from $O(1)$ to $O(N \cdot d)$. Following these passes, each sequence is expanded over the vocabulary $\mathcal{V}_f$, yielding a large candidate set for ranking. Their scores (e.g., CFS) are computed incrementally by updating the accumulated log-probabilities with the log-probabilities obtained from the same forward passes, requiring no additional model invocations. These candidates are then ranked to retain the top sequences; using partial sorting (e.g., top-k selection), this incurs a cost of $O(N \cdot d \cdot |\mathcal{V}_f| \log(N \cdot d))$ per step. In Stage II, the computational overhead is further optimized, reducing the requirement to $O(Q)$ forward passes and a ranking cost of $O(Q \cdot |\mathcal{V}_f| \log Q)$ per step. TRACE makes this search tractable by restricting the branching factor from the full vocabulary $\mathcal{V}$ to a highly compact candidate set $\mathcal{V}_f$. We acknowledge the computational overhead associated with the reconstruction phase; however, within an offline threat model based on static snapshots, prioritizing extraction fidelity over real-time latency represents a reasonable and practical compromise.

D Experiments

D.1 Datasets

We evaluate TRACE on various representative benchmarks for LLM unlearning, including TOFU, MUSE, WMDP, and RWKU.

TOFU [36]. We use the TOFU (Task Of Fictitious Unlearning) benchmark, which consists of synthetic author profiles paired with question–answer data. The dataset contains 200 fictitious authors, each associated with 20 question–answer pairs, resulting in 4,000 samples in total. We construct the training data directly from this dataset, randomly sampling a subset as the forget set, with the remaining samples serving as the retain set.

MUSE [37]. We use the MUSE-News subset from the MUSE (Machine Unlearning Six-way Evaluation for language models) benchmark, which is constructed from real-world news corpora. The training data consists of approximately 6,300 sentences randomly sampled from the dataset. The forget data is randomly sampling from the training data.

WMDP [38]. We use the bio-retain-corpus from the WMDP (Weapons of Mass Destruction Proxy) benchmark, which consists of biomedical text designed to represent safe, retainable knowledge. The training data is constructed by randomly sampling 1,000 segments from the corpus, each truncated to a maximum length of 512 tokens. The forget data is randomly sampling from the training data.

RWKU [44]. We use approximately 6,200 samples from the RWKU (Real-World Knowledge Unlearning) benchmark, which are derived from real-world knowledge about public entities (e.g., well-known individuals). These samples consist of factual descriptions, such as biographical details and related facts. The forget set consists of text segments associated with a specific entity (e.g., a public figure), and the prefixes are extracted directly from these segments, typically consisting of the

Algorithm 1 TRACE

Input: pre-unlearning model θ^*, post-unlearning model θ_u, expansion factor d, completion search width Q, number of samples to reconstruct N, similarity threshold τ, maximum generation length $L_{\max}$
Output: reconstructed set $\mathcal{S}$

$\mathcal{V}_f \leftarrow \textsc{ExtractTokens}(\theta^*, \theta_u)$ ▷ defined in Eq. (1)
$\mathcal{X}^{(0)} \leftarrow \{v \in \mathcal{V}_f \mid v \text{ satisfies linguistic constraints}\}$ ▷ valid initial tokens
$\mathcal{S} \leftarrow \emptyset$ ▷ reconstructed sequences
Stage I: Broad Prefix Exploration
for $r = 1, 2, \ldots, L_{\max}$ **do**
  $r^* \leftarrow r$
  $\mathcal{C}^{(r)} = \{\, x \oplus v \mid x \in \mathcal{X}^{(r-1)},\ v \in \mathcal{V}_f \,\}$
  $\mathcal{X}^{(r)} \leftarrow \mathrm{Top}_{N\times d}\big(\mathcal{C}^{(r)}; \mathrm{CFS}(\cdot;\alpha_s,\beta)\big) \cup \mathrm{Top}_{N\times d}\big(\mathcal{C}^{(r)}; \mathrm{CFS}(\cdot;\alpha_l,\beta)\big)$
  $\mathcal{P}^{(r)} \leftarrow \textsc{Cluster}(\mathcal{X}^{(r)}; \tau)$ ▷ defined in Eq. (6)
  if $|\mathcal{P}^{(r)}| \le N$ **then**
    break
  end if
end for
Stage II: Completion from Recorded Prefixes
$\mathcal{U} \leftarrow \mathcal{X}^{(r^*)}$
if $r^* = L_{\max}$ **then**
  return $\mathcal{U}$
end if
while $\mathcal{U} \neq \emptyset$ **do**
  Initialize $\mathcal{X} \leftarrow \mathcal{U}$
  $\texttt{stop}_s, \texttt{stop}_l \leftarrow$ False
  while not ($\texttt{stop}_s$ and $\texttt{stop}_l$) **do**
    $\mathcal{C} \leftarrow \{x \oplus v \mid x \in \mathcal{X}, v \in \mathcal{V}_f\}$
    if not $\texttt{stop}_s$ and $\mathrm{IsComplete}(\mathrm{Top}_1(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_s,\beta)))$ **then** ▷ see Sec. B.3
      $\mathcal{S} \leftarrow \mathcal{S} \cup \mathrm{Top}_1(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_s,\beta))$
      $\texttt{stop}_s \leftarrow$ True
    end if
    if not $\texttt{stop}_l$ and $\mathrm{IsComplete}(\mathrm{Top}_1(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_l,\beta)))$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \mathrm{Top}_1(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_l,\beta))$
      $\texttt{stop}_l \leftarrow$ True
    end if
    if $\texttt{stop}_s$ **then**
      $\mathcal{X} \leftarrow \mathrm{Top}_Q(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_l,\beta))$
    else if $\texttt{stop}_l$ **then**
      $\mathcal{X} \leftarrow \mathrm{Top}_Q(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_s,\beta))$
    else
      $\mathcal{X} \leftarrow \mathrm{Top}_Q(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_s,\beta)) \cup \mathrm{Top}_Q(\mathcal{C}; \mathrm{CFS}(\cdot;\alpha_l,\beta))$
    end if
  end while
  $\mathcal{U} \leftarrow \{x \in \mathcal{U} \mid \max_{s\in\mathcal{S}} \text{ROUGE-L}_{recall}(x, s) < \tau\}$
end while
return $\mathcal{S}$

entity name followed by contextual words (e.g., "John was", "John began", "John has" when using the first two words as the prefix).

D.2 Metrics

Follow [19], three metrics are adopted for evaluation, including ROUGE-L recall (R), and average extraction success rate (A-ESR$_\kappa$). A-ESR$_\kappa$ measures the percentage of ground-truth samples that can be successfully reconstructed under a similarity threshold κ. Specifically, a ground-truth sample

is considered successfully reconstructed if there exists at least one reconstructed sample whose ROUGE-L recall exceeds κ. Formally, given the ground-truth set $\mathcal{X}_{GT}$ and reconstructed set $\mathcal{S}$, we define:

$$\text{A-ESR}_\kappa(\mathcal{X}_{GT}, \mathcal{S}) = \frac{1}{|\mathcal{X}_{GT}|} \sum_{x \in \mathcal{X}_{GT}} \mathbf{1}\left[\max_{s \in \mathcal{S}} \text{ROUGE-L(R)}(x, s) \geq \kappa\right] \times 100\%, \tag{8}$$

where $\mathbf{1}[\cdot]$ is the indicator function. We report A-ESR$_{1.0}$ (exact match) and A-ESR$_{0.9}$ (approximate match) by default.

D.3 Implementation Details

Experiments are conducted on a server equipped with 8 NVIDIA A100 GPUs.

Training and Unlearning. All pre-unlearning models are trained on the corresponding training data for 5 epochs with a learning rate of 1×10^{-5}. For unlearning, we determine the number of unlearning epochs based on the perplexity of a retrained model (i.e., a model trained from scratch without the forget set). Specifically, the unlearning process is performed for 2 to 6 epochs, and we select the checkpoint whose perplexity on both the unlearned and retained data is closest to that of the retrained model.

Candidate Tokens Extraction. For most models, the candidate token set V_f can be directly obtained using Eq. (1), as unlearning typically induces sparse embedding changes, and only a small subset of tokens exhibits non-zero differences. However, for models that tie the input embedding matrix to the output projection layer (e.g., Qwen), unlearning causes updates to propagate across the entire vocabulary in the embedding space. This makes the thresholding criterion in Eq. (1) less effective, as V_f would otherwise include nearly the full vocabulary.

Despite this, we observe that tokens involved in the unlearning process tend to exhibit significantly larger embedding changes compared to unrelated tokens. Based on this observation, for models with tied embeddings, we construct V_f by selecting the top-m tokens with the largest embedding difference norms $\Delta(v)$. In our experiments, we use $m = 6000$ for Qwen2.5-3B. As shown in Table 1, TRACE achieves high-quality reconstruction on Qwen2.5-3B, even though V_f contains a large number of unrelated tokens.

Hyperparameter Settings. We set the CFS parameters $(\alpha_s, \alpha_l, \beta) = (2.0, 6.0, 0.5)$ for TOFU, MUSE, and RWKU, and $(1.5, 6.0, 0.3)$ for WMDP. The expansion factor d is set to 10 for Qwen2.5-3B and StableLM-3B, and 5 for other models on TOFU and MUSE. The completion search width Q is set to 10 for TOFU and MUSE, 20 for RWKU, and 40 for WMDP. We use a similarity threshold $\tau = 0.7$ for TOFU and $\tau = 0.6$ for all other datasets. The maximum generation length is set to $L_{\max} = 150$ for TOFU and MUSE, 400 for RWKU, and 512 for WMDP, reflecting dataset-specific sequence length ranges.

D.4 Discussion and Comparison with Reconstruction Baselines

Training Data Reconstruction. Existing training data reconstruction attacks are not directly applicable to our setting in general. Gradient-based reconstruction methods [26, 27, 28, 29] assume access to gradients produced during training, while other attacks exploit internal representations, such as embeddings or intermediate activations [30, 31]. Such data-dependent information is unavailable under our threat model, where the attacker has access to the pre- and post-unlearning models but not to gradients or intermediate activations associated with the unlearned samples. Among these methods, FILM [27] provides a relevant comparison, as it reconstructs training text in llms from gradients in federated learning. Importantly, FILM performs approximate reconstruction from the gradients rather than requiring exact gradients. We therefore adapt FILM to our setting by treating the parameter difference between the pre- and post-unlearning models as a surrogate gradient signal for the unlearned data and applying its reconstruction procedure accordingly. For a controlled comparison, we configure FILM to generate the same number of samples as TRACE. We compare TRACE, PPL-only, and FILM on TOFU and MUSE using LLaMA2-7B under GA and GD.

As shown in Table 11, TRACE substantially outperforms FILM. Notably, under the same number of generated samples, PPL-only baseline also achieves stronger reconstruction performance and

reconstructs more unlearned samples than FILM, partly due to its adoption of TRACE's two-stage architecture. These results establish PPL-only as a strong reconstruction baseline in our main experiments.

Table 11: Reconstruction performance of TRACE, PPL-only, and FILM.

		GA			GD		
Dataset	Metric	TRACE	PPL-only	FILM	TRACE	PPL-only	FILM
TOFU	ROUGE-L(R)	**0.9799**	0.6017	0.4355	**0.9695**	0.2946	0.2393
	A-ESR$_{1.0}$ (%)	**95.00**	28.75	1.25	**92.50**	7.50	0.00
	A-ESR$_{0.9}$ (%)	**97.50**	36.25	5.00	**96.25**	10.00	0.00
MUSE	ROUGE-L(R)	**0.9457**	0.3915	0.2625	**0.9332**	0.1654	0.1968
	A-ESR$_{1.0}$ (%)	**93.75**	37.50	10.00	**92.50**	13.75	7.50
	A-ESR$_{0.9}$ (%)	**93.75**	37.50	15.00	**92.50**	13.75	10.00

Unlearning Inversion Attacks. Among prior work on unlearning inversion [18, 32, 19, 33], UBNF [19] introduced unlearned text reconstruction for LLMs by exploiting output likelihood differences between the original and unlearned models. However, UBNF requires a known prefix of the target sample to initiate reconstruction. In contrast, TRACE is the first method to reconstruct unlearned samples from LLMs without target prefixes. To further compare their reconstruction capability under the same prior information, we additionally evaluate TRACE and UBNF with identical target-prefix information. As evaluated in Section 5.3, when both methods are provided with the same ground-truth prefixes, TRACE consistently achieves higher reconstruction performance and recovers a larger fraction of the remaining target content.

D.5 Reconstruction under Exact Retraining

Our main experiments focus on approximate unlearning methods, which modify a model to remove the influence of selected training samples. We further investigate whether the reconstruction risk exposed by TRACE persists under exact unlearning through retraining. Specifically, we train the pre-unlearning model on the original MUSE training set and the retrained model on the retained data after removing the target samples, using 4 training epochs for both models. We then apply TRACE to the two models to reconstruct the unlearned samples. For comparison, we report the corresponding white-box and black-box results under the approximate unlearning method NPO with the same unlearning batch size of 8.

Table 12: TRACE under extract unlearning (Retrain).

Metric	Npo(white)	Npo(black)	Retrain
ROUGE-L(R)	0.6566	0.3419	0.2068
A-ESR$_{1.0}$	51.25	27.50	5.00
A-ESR$_{0.9}$	55.00	27.50	5.00

As shown in Table 12, exact retraining substantially reduces reconstruction risk compared with approximate unlearning. Nevertheless, TRACE can still exactly reconstruct a small fraction of the unlearned samples. The markedly higher reconstruction performance under approximate unlearning suggests that its targeted model updates leave stronger exploitable traces of the unlearned data.

E Visualization

E.1 Multi-sample reconstruction without prior knowledge of unlearned data

We provide an example of multi-sample reconstruction without prior knowledge of the target data using LLaMA2-7B on MUSE under NPO, with an unlearning batch size of 32. For each unlearned sample, we present the reconstructed sequence with the highest similarity to the ground truth, as shown in Table 13.

Note on Data Usage. To further mitigate privacy concerns, we anonymize all personal names in the visualization by replacing them with placeholders (e.g., "[Name]").

Table 13: Comparison of reconstructed multiple samples by TRACE and PPL-only baseline on MUSE dataset. Red text indicates lexical overlap with the ground truth.

#	Unlearned Data (GT)	TRACE (Ours)	PPL-only (Baseline)
1	Supermarkets are expected to be allowed to select which items they would cap and only take part in the initiative, modelled on a similar agreement in France, on a voluntary basis.	[Name], 17, pleaded guilty to first-degree murder in April for his role in the 2 November 2021 killing of [Name], a Spanish teacher.	"These businesses want to be part of a sustainable future, but they can't be part of a sustainable future on their own.
2	In a separate survey it found firms have seen an average 53% drop in sales compared with pre-war 2021. It reported that larger companies have suffered more disruption that smaller firms.	In a separate survey it found firms have seen an average 53% drop in sales compared with pre-war 2021. It reported that larger companies have suffered more disruption that smaller firms.	In a separate survey it found firms have seen an average 53% drop in sales compared with pre-war 2021. It reported that larger companies have suffered more disruption that smaller firms.
3	It feels like a process which has now passed its best before date.	It feels like a process which has now passed its best before date.	Scotland is a progressive country, Ms [Name] said, and the country's people want the government to be seen to be taking action on problems like drug deaths.
4	[Name] said his son was like "a glue" for lots of different friend-ship groups	[Name] said his son was like "a glue" for lots of different groups [Name] said his son was like "a. Webber said his son was like "a glue" for lots of different groups.	However, the BBC said it was "very different" to the one in the 2000s.
5	Being in default renders Kemble Water effectively insolvent but that will not affect the underlying utility business that supplies water and waste services to 1 in 4 people in the UK.	Being in default renders Kemble Water effectively insolvent but that will not affect the underlying utility business that supplies water and waste services to 1 in 4 people in the UK.	In a statement to the BBC, the Scottish government said that road safety and investing in infrastruc-ture were high priorities.
6	She thinks it is alarming that Chi-nese authorities feel comfortable targeting foreign citizens and attempting to dictate their work.	She thinks it is alarming that Chi-nese authorities feel comfortable targeting foreign citizens and at-tempting to dictate their work.	She said: "When I found out that I was pregnant again I was - I didn't want to be, I was sad about it.
7	[Name], 17, pleaded guilty to first-degree murder in April for his role in the 2 November 2021 killing of [Name], a Spanish teacher.	[Name], 17, pleaded guilty to first-degree murder in April for his role in the 2 November 2021 killing of [Name], a Spanish teacher.	However, the BBC said it was "very different" to the one in the 2000s.
8	The government argued that any-one identified as a potential victim of modern slavery would be re-turned home or to another "safe country away from those who have trafficked them".	The government argued that any-one identified as a potential victim of modern slavery would be re-turned home or to another "safe country away from those who have trafficked them".	Scotland is a progressive country, Ms [Name] said, and the country's people want the government to be seen to be taking action on problems like drug deaths.
9	"With owners, I think people need to start realising that they have responsibilities.	"With owners, I think people need to start realising that they have responsibilities.	In a separate survey it found firms have seen an average 53% drop in sales compared with pre-war 2021. It reported that larger companies have suffered more disruption that smaller firms.
10	After her death at the age of 56, here's a closer look at how she recorded her most famous song.	After her death at the age of 56, here's a closer look at how she recorded her most famous song.	Ms [Name] had insisted at the time that she was "not concerned" about the party's finances.

Continued on next page

#	Unlearned Data (GT)	TRACE (Ours)	PPL-only (Baseline)
11	With more bad weather forecast for Sunday, England are left hoping for enough time to force the win that would level the series at 2-2 and turn the fifth Test at The Oval into a decider.	With more bad weather forecast for Sunday, England are left hoping for enough time to force the win that would level the series at 2-2 and turn the fifth Test at The Oval into a decider.	In a statement to the BBC, the Scottish government said that road safety and investing in infrastructure were high priorities.
12	In the BBC programme Surviving Aberfan, survivor [Name] recounted how his life was saved by Mr [Name].	In the BBC programme Surviving Aberfan, survivor[Name] recounted how his life was saved by Mr [Name].	However, the BBC said it was "very different" to the one in the 2000s.
13	Ms [Name] had insisted at the time that she was "not concerned" about the party's finances.	Ms [Name] had insisted at the time that she was "not concerned" about the party's finances.	Ms [Name] had insisted at the time that she was "not concerned" about the party's finances.
14	[Name] have signed France striker [Name] from RB Leipzig for £52m.	[Name] have signed France striker [Name] from RB Leipzig for £52m.	[Name] have signed France striker [Name] from RB Leipzig for £52m.
15	"[Name] was a caring daughter, wife, sister, mother and friend. If love and compassion could make a person immortal, she would have lived forever.	"[Name] was a caring daughter, wife, sister, mother and friend. If love and compassion could make a person immortal, she would have lived forever.	The jury heard that if her son had not been found when he was, it was estimated he would have been in the water for about 15 or 16 hours.
16	The symptoms depend on the individual, which drug they were taking, and for how long. Many patients can stop taking antidepressants without experiencing any problems.	He hailed the late star as a "great black female" and a "positive image for any woman", having overcome "hard times" in her personal life.	Scotland is a progressive country, Ms [Name] said, and the country's people want the government to be seen to be taking action on problems like drug deaths.
17	Based on the information she provided the British Pregnancy Advisory Service (BPAS), she was sent the tablets because it was estimated she was seven weeks pregnant.	In 2019, Ms [Name] had insisted she was "not concerned" about the party's finances, and said it was "in the black" at the end of 2019.	Ms [Name] had insisted at the time that she was "not concerned" about the party's finances.
18	The wait time at one stage was estimated to be more than 24 hours; and the queue snaked from Westminster Hall, down along the River Thames and stretched south for almost seven miles (11km).	Officials said the government would maintain a range of contingency measures to "enable detection, and swift and proportionate action, for potential new harmful variants" should the need arise.	However, the BBC said it was "very different" to the one in the 2000s.
19	[Name], from Staffordshire, already had three sons before she became pregnant again in 2019.	[Name], from Staffordshire, already had three sons before she became pregnant again in 2019.	She said: "When I found out that I was pregnant again I was - I didn't want to be, I was sad about it.
20	The Scottish government said that road safety and investing in infrastructure were high priorities.	The Scottish government said that road safety and investing in infrastructure were high priorities.	The Scottish government said that road safety and investing in infrastructure were high priorities.
21	"When I read they had heard banging noises I was feeling hopeful that maybe it was coming from the submersible. But then yesterday was the worst when I heard that he didn't make it, that they all died," she said.	"When I read they had heard banging noises I was feeling hopeful that maybe it was coming from the submersible. But then yesterday was the worst when I heard that he didn't make it, that they all died," she said.	She said: "When I found out that I was pregnant again I was - I didn't want to be, I was sad about it.
22	Since then, tensions between the Sunni- and Shia-led neighbours have often been high, with each regarding the other as a threatening power seeking regional dominance.	Since then, tensions between the Sunni- and Shia-led neighbours have often been high, with each regarding the other as a threatening power seeking regional dominance.	Since then, tensions between the Sunni- and Shia-led neighbours have often been high, with each regarding the other as a threatening power seeking regional dominance.

Continued on next page

#	Unlearned Data (GT)	TRACE (Ours)	PPL-only (Baseline)
23	None 'An unstoppable juggernaut' - where will Man City dominance end?	Since then, tensions between the Sunni- and Shia-led neighbours have often been high, with each regarding the other as a threatening power seeking regional dominance.	Since then, tensions between the Sunni- and Shia-led neighbours have often been high, with each regarding the other as a threatening power seeking regional dominance.
24	Officials said the government would maintain a range of contingency measures to "enable detection, and swift and proportionate action, for potential new harmful variants" should the need arise.	Officials said the government would maintain a range of contingency measures to "enable detection, and swift and proportionate action, for potential new harmful variants" should the need arise.	"We are investing over £142m, including £42m in 2023-2024, in a range of measures aimed at reducing the number of patients having to wait over 12 hours, or 72 hours from the time they need to be seen, to a target of 10,000 in 2024-2025," she said.
25	However, the advocate told the jury that they had to put such feelings aside.	However, the advocate told the jury that they had to put such feelings aside.	However, the BBC said it was "very different" to the one in the 2000s.
26	He hailed the late star as a "great black female" and a "positive image for any woman", having overcome "hard times" in her personal life.	He hailed the late star as a "great black female" and a "positive image for any woman", having overcome "hard times" in her personal life.	The Scottish government said that road safety and investing in infrastructure were high priorities.
27	"These events sadden the England cricket teams, and we are thinking about those affected at this harrowing time. As a gesture of respect, we will honour them by wearing black armbands."	"These events sadden the England cricket teams, and we are thinking about those affected at this harrowing time. As a gesture of respect, we will honour them by wearing black armbands."	"We are investing over £142m, including £42m in 2023-2024, in a range of measures aimed at reducing the number of patients having to wait over 12 hours, or 72 hours from the time they need to be seen, to a target of 10,000 in 2024-2025," she said.
28	It was first obtained by CNN, but the BBC's US partner CBS also has the clip.	It was first obtained by CNN, but the BBC's US partner CBS also has the clip.	However, the BBC said it was "very different" to the one in the 2000s.
29	Detective Chief Superintendent [Name] of the City of London Police said: "We approached the CBI following media reports and our investigations are at a very early stage. It would not be appropriate to comment any further at this time."	Detective Chief Superintendent [Name] of the City of London Police said: "We approached the CBI following media reports and our investigations are at a very early stage. It would not be appropriate to comment any further at this time."	In the 1970s, when he was in his 20s, he was involved in a number of businesses, including as a partner in a St Christopher's Christopher's) and was a "significant investor" in a number of other companies, one of which was a "mining business" in the early 1990s. It was at this time that he also became a "significant investor" in a number of other companies, one of which was a "mining business" in the early 1990s.
30	Under his direction, the CFPB increased its scrutiny of banks and their customer fees in 2022, soliciting complaints from consumers about practices such as overdraft fees.	Under his direction, the CFPB increased its scrutiny of banks and their customer fees in 2022, soliciting complaints from consumers about practices such as overdraft fees.	The Scottish government said that road safety and investing in infrastructure were high priorities.
31	[Name] said no-one else is believed to be involved. Mr [Name]' death is not considered suspicious.	[Name] said no-one else is believed to be involved. Mr [Name]' death is not considered suspicious.	[Name] said no-one else is believed to be involved. Mr [Name]' death is not considered suspicious.

Continued on next page

#	Unlearned Data (GT)	TRACE (Ours)	PPL-only (Baseline)
32	The Commonwealth Games Federation (CGF) says it is "considering multiple proposals" to host the event in 2026, amid continued uncertainty over its long-term future. Last week, Singapore joined Malaysia in ruling out a bid, after the Australian state of Victoria withdrew as the original host in July because of rising costs. However, in its first public statement since the news, the CGF said "significant progress had been made", and it was "excited by the early concepts, which aim to reset and reframe the Games". It claimed it was working with a number of Commonwealth Games Associations (CGAs) over "further detailed assessments", with an aim to announce a host in May. It did not identify which countries were involved, saying the CGAs had asked to keep the proposals confidential. "The CGF continues to accelerate work to refresh the Games… including exploring innovative new concepts," it said. "Importantly, we have been focusing on how we can transform the Games to a collaborative and truly sustainable model, minimising costs and reducing its environmental footprint, whilst increasing social impact."	In 2020 Ms [Name] said she was "not concerned" about the party's finances, saying that "in 2021-22 we were £666,000 in the black; in 2022-23 we were £266,000 in the black; and in 2023-24 we expected to be £166,000 in the black." She said the party was "in the black" for the first time since 2019-2020.	In the 1970s, she became one of the BBC's BBC's first female regional news presenters, working in London, then in 1994 she became the first woman to read the news on the BBC's UK regional news programme.

E.2 Comparison with Prefix-Assisted UBNF.

We present a comparison between TRACE and prefix-assisted UBNF on a representative unlearned sample from the WMDP dataset using LLaMA2-7B, under NPO (see Table 14).

Table 14: An example of reconstruction on WMDP dataset. TRACE operates without prior knowledge, while UBNF is provided with a prefix (10 words). Red text indicates lexical overlap with the ground truth.

Unlearned text	A viral zoonos is called monkeypox is endemic to some regions of Africa. Its primary symptoms, like those of other illnesses brought on by pox viruses, are fever and skin lesions. Unfortunately, a small percentage of individuals may experience severe, multi-systemic illness that quite frequently can be deadly. Due to a rapid outbreak of cases discovered outside of its endemic range, monkeypox has lately attracted attention and worry on a global scale. The smallpox (variola) virus and the monkeypox virus are both enveloped double-stranded DNA viruses belonging to the Poxviridae family. In the Democratic Republic of the Congo, monkeypox was first found in a human in 1970. Several epidemics have occurred since then, primarily impacting African nations. Subsequently, instances outside of Africa were also documented, and the illness gained relevance for public health. The World Health Organization (WHO) has recently confirmed an atypical outbreak of monkeypox from multiple non-endemic countries with increasing numbers of cases reported almost daily. The WHO has declared monkeypox to be a public health emergency of global concern. Monkeypox outbreaks have been documented in 75 nations so far, and they are rapidly expanding around the globe. Typically, monkeypox is a self-limiting illness with symptoms that last between two and four weeks. Children are more likely to experience severe symptoms, which are connected to the level of viral exposure, the patient's condition, and the kind of problems. The results might be worse if immunological deficits are present. Although smallpox immunization has proved to be protective in the past, people in the age group of 40 to 50 may now be more vulnerable to monkeypox due to the worldwide discontinuation of smallpox vaccine programs after the illness was eradicated. Monkeypox complications can include secondary infections, bronchopneumonia, sepsis, encephalitis, and corneal infections with subsequent vision loss. It is uncertain how widespread an asymptomatic infection could be. Whereas this skin lesions and rashes of monkeypox frequently res
Input for TRACE	[redacted]
TRACE (Ours)	A viral zoonos is called monkeypox is endemic to some regions of Africa. Its primary symptoms, like those of other illnesses brought on by pox viruses, are fever and skin lesions. Unfortunately, a small percentage of individuals may experience severe, multi-systemic illness that quite frequently can be deadly. Due to a rapid outbreak of cases discovered outside of its endemic range, monkeypox has lately attracted attention and worry on a global scale. The smallpox (variola) virus and the monkeypox virus are both enveloped double-stranded DNA viruses belonging to the Poxviridae family. In the Democratic Republic of the Congo, monkeypox was first found in a human in 1970. Several epidemics have occurred since then, primarily impacting African nations. Subsequently, instances outside of Africa were also documented, and the illness gained relevance for public health. The World Health Organization (WHO) has recently confirmed an atypical outbreak of monkeypox from multiple non-endemic countries with increasing numbers of cases reported almost daily. The WHO has declared monkeypox to be a public health emergency of global concern. Monkeypox outbreaks have been documented in 75 nations so far, and they are rapidly expanding around the globe. Typically, monkeypox is a self-limiting illness with symptoms that last between two and four weeks. Children are more likely to experience severe symptoms, which are connected to the level of viral exposure, the patient's condition, and the kind of problems. Although smallpox immunization has proved to be protective in the past, people in the age group of 40 to 50 may now be more vulnerable to monkeypox due to the worldwide discontinuation of smallpox vaccine programs after the illness was eradicated. Monkeypox complications can include secondary infections, bronchopneumonia, sepsis, encephalitis, and corneal infections with subsequent vision loss. Although smallpox vaccine has proved to be protective in the past, people in the age group of 40 to 50 may now be more vulnerable to monkeypox due to discontinuation of smallpox
Input for UNBF	**A viral zoonosis called monkeypox is endemic to some regions** [redacted]

Continued on next page

Continued from previous page	
UNBF (Baseline)	**A viral zoonosis called monkeypox is endemic to some regions** of Africa. Its primary symptoms, like those of other illnesses brought on by pox viruses, are fever and skin lesions. Unfortunately, a small percentage of individuals may experience severe, multi-systemic illness that frequently requires hospitalization and, in some cases, intensive care unit support. Due to complexities in identifying monkeypox virusal strains and the ongoing evolution of the pandemic, effective treatments have yet to be approved. Antiviral drug treatment of monkeypox has primarily relied on small molecule therapies with established safety profiles for the treatment of other viral infections. These include the use of brincidofovir, a broad-spectrum DNA polyomavirus and poxvirus antiviral, and the nucleoside analogs, favipiravir and tecovirimat, which have shown potential in treating monkeypox in animal models. However, the efficacy of these drugs in clinical settings has yet to be demonstrated. In addition to the use of approved drugs, there is a pressing need to identify small molecules with specific activity against monkeypox virus. Natural product-derived small molecule libraries offer a rich source of potential leads with this type of specificity. In particular, compounds that inhibit poxvirus assembly and egress have been identified from the lanosterol library of Fasciola gigantica, a parasitic flatworm. The most potent of these, a compound designated Fas2_32, was found to have 100-fold selectivity for the mammalian host over the sheep barber pole assembly sites used by the sheep infecting strain of Old World orthopoxvirus, vaccinia virus. In addition, Fas2_32 was found to be active against the highly related variola virus, the causative agent of smallpox, with an EC50 of 1.3 03bcM. The epsilon toxin (ETX) of monkeypox is a 66 kDa envelope gly

F Limitations

While our results demonstrate that approximate unlearning in LLMs can leave recoverable traces in model parameters, our study is subject to several limitations.

First, our threat model assumes access to both pre- and post-unlearning model snapshots with white-box parameter access. While this setting is realistic in scenarios such as model auditing, rollback, or internal MLaaS pipelines where multiple checkpoints are maintained, it may not apply to strictly closed-source deployments where such access is restricted. Second, our reconstruction framework relies on parameter differences in the embedding layer to identify candidate tokens. Its effectiveness may vary under future unlearning techniques specifically designed to obscure such signals. Finally, TRACE focuses on reconstructing textual data that has been explicitly targeted by the unlearning process. The reconstruction quality is influenced by factors such as model memorization strength, and unlearning batch size.

Despite these limitations, we believe our setting captures realistic deployment practices where multiple model versions are maintained, and our findings provide actionable insights for improving the safety and reliability of machine unlearning systems.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   The abstract and introduction clearly present the main claim of this paper: that approximate unlearning in large language models can leave recoverable traces that enable reconstruction of unlearned samples. This claim is supported by the TRACE framework introduced in Sec. 4. The effectiveness and generality of this approach are validated through extensive experiments in Sec. 5 across multiple datasets, model architectures, and representative unlearning methods. Furthermore, we explicitly define the threat model and assumptions in Sec. 3, ensuring that the scope and limitations of our claims are clearly articulated and aligned with practical deployment scenarios.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations in Appendix Sec. F

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not involve formal theoretical assumptions or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main setup is described in Sec. 5.1, with additional implementation details provided in Appendix Sec. D.3, ensuring reproducibility of the results relevant to the core claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not use any private data, and all datasets used in this work are publicly available. We provide detailed implementation and experimental settings in the paper and appendix to support reproducibility, along with the pseudocode of our method. While the code is not publicly available at submission time, we commit to release it upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main setup is described in Sec. 5.1, with additional details provided in Appendix Sec. D.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: As shown in Table 1, we evaluate multiple models, datasets, and unlearning methods. For each setting, results are averaged over 10 randomly sampled unlearning runs. We observe that the variation in ROUGE-L(R) across runs is relatively small (standard deviation < 0.05), while the performance gap between TRACE and the baseline remains significantly larger. For clarity, we report only the mean values in Table 1 and error bars are omitted.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information available in Appendix Sec. D.3

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and ensured that the research complies with all relevant ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the broader impacts in Appendix Sec. A

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not center on releasing any new datasets or models, and thus no additional safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use multiple benchmark datasets, all of which are properly cited in Sec. 5.1, with detailed descriptions provided in Appendix Sec. D.1.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: In this work, large language models are the target of analysis rather than a component of the proposed method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.