MODE-AVOIDING UNLEARNING VIA NEGATIVE-ORDER RÉNYI DIVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on continual learning in Large Language Models (LLMs) shows that on-policy finetuning is implicitly biased toward the solution of minimal KL divergence from the base model, allowing reverse-KL-driven objectives to learn new knowledge while largely preserving prior capabilities. We argue that unlearning can be viewed as the inverse analogue of continual learning: rather than adding a mode while preserving the rest of the distribution, it must remove a specific forget mode while leaving the rest of the distribution untouched. Existing unlearning methods, however, typically rely on off-policy objectives that push the model toward an idealized unlearning target, leading to utility collapse. Since KL divergence is itself the $\alpha \to 1$ limit of the Rényi divergence family, and every positive-order Rényi divergence only attracts a model toward a target distribution, we introduce MARU (Mode-Avoiding Rényi Unlearning), a method that uses the negative-order Rényi divergence as a natural mode-avoiding objective for unlearning. Anchored at the pre-unlearning policy, this yields an objective whose geometry implicitly preserves model utility without requiring an explicit retain set. Extensive experiments on the TOFU and MUSE benchmarks show that MARU achieves state-of-the-art performance on the hardest forget split (10%) at 1B scale, while delivering up to $31.5\times$ better resistance to membership inference attacks than comparable distillation-based methods, though these gains are weaker on MUSE. These results suggest that reframing unlearning through divergence geometry, rather than heuristic gradient objectives, offers a principled and effective path toward safe, controllable and utility-preserving unlearning in LLMs.

1 INTRODUCTION

Large Language Models (LLMs) owe much of their broad capability to training on massive, web-scale corpora. Yet this same scale makes it difficult to control what knowledge and skills the models have absorbed. Prior work has shown that LLMs can memorize and later reproduce personal information, copyrighted text, and other sensitive content present in their training data (Carlini et al., 2021). These risks have taken on new urgency as data protection regulations, including the GDPR's "right to be forgotten" (European Union, 2016), the California Consumer Privacy Act (CCPA) (Pardau, 2018), and the EU AI Act (European Union, 2023), increasingly require that specific information be removed from deployed systems upon request. Since retraining an LLM from scratch to remove such content is computationally prohibitive, Machine Unlearning (MU) has emerged as a practical alternative, aiming to selectively erase targeted knowledge from a trained model while preserving its general capabilities.

Existing approaches to LLM unlearning are typically framed as a multi-objective problem, pairing a forget loss that suppresses targeted generations with an explicit retain regularizer that preserves utility on retain data (Liu et al., 2025; Dorna et al., 2026). Within this landscape, gradient-based objectives such as Gradient Ascent (GA) and Gradient Difference (GD) directly penalize likelihood on the forget set; preference-style losses such as Negative Preference Optimization (NPO) and SimNPO reformulate forgetting as a preference-alignment problem that discourages specific outputs (Zhang et al., 2024; Fan et al., 2026); and self-distillation-based objectives such as UNDIAL (Dong et al., 2025) and DUET (Zhong et al., 2026) transfer knowledge from a teacher model that guides the information-removal process. Despite this diversity, these methods share two limitations: the forget objective is typically chosen heuristically rather than derived from any account of why certain train-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

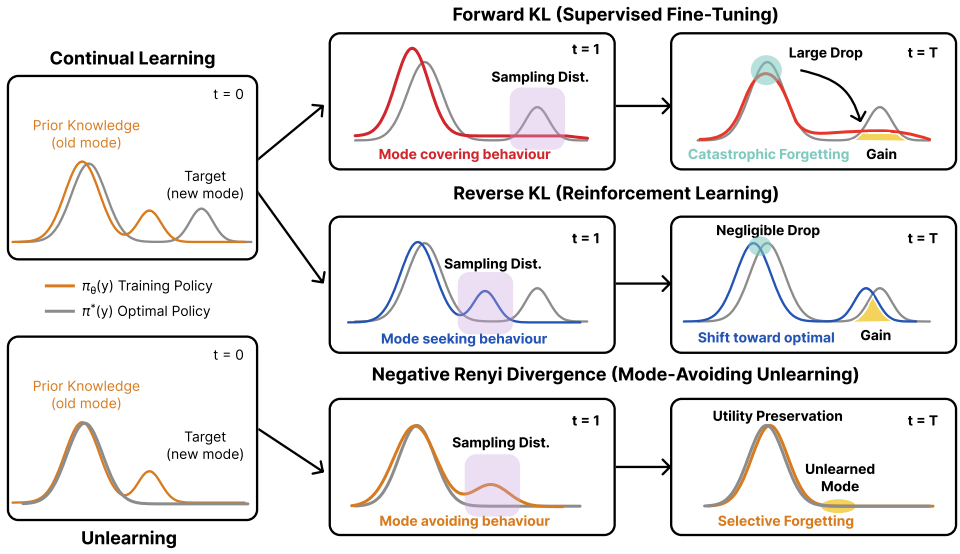


Figure 1: Self-distillation induces reverse-KL, mode-seeking dynamics: unlike forward-KL (SFT), which stretches to cover both old and new modes and drops mass from prior knowledge, reverse-KL (RL) fitting shifts the existing mode toward the target while preserving its shape (Chen et al., 2026). We extend this same geometry to unlearning, where instead of moving toward a new mode the objective moves away from an existing one without collateral distortion elsewhere.

ing dynamics preserve capability while others erode it, and utility is protected only by pairing the forget loss with an explicit retain regularizer, which adds retain data, hyperparameters, and compute.

A recent line of work originating from the continual learning for LLMs literature offers a principled account of the utility preservation dynamics. On-policy fine-tuning has been shown to forget substantially less than supervised fine-tuning (Shenfeld et al., 2026b), an effect attributed to on-policy sampling being implicitly biased toward solutions of minimal Kullback-Leibler (KL) divergence from the base model. Supervised fine-tuning is characterized as mode-covering: minimizing forward KL forces the model to place mass wherever the target data does, which means displacing previously-acquired knowledge and causing catastrophic forgetting. The reverse KL objective, on the other hand, is mode-seeking, meaning that updates concentrate on the target's high-density regions while leaving the rest of the distribution intact. We carry this insight over to unlearning. To the best of our knowledge we are the first to argue that:

> Unlearning can be viewed as the **inverse** analogue of continual learning: rather than adding a mode while preserving the rest, we must remove a specific forget mode while leaving the rest of the distribution untouched.

Realizing this inversion requires a divergence that is *repulsive* rather than attractive, and none of the standard positive-order divergences qualify. The KL divergence is the $\alpha \to 1$ limit of the Rényi divergence family, and every positive-order Rényi divergence is attractive, minimized only when the model matches its target distribution. This means that there is no way to unlearn within the $\alpha > 0$ regime. We instead turn to the negative-order Rényi divergence, where the functional changes sign and becomes repulsive (Li & Turner, 2016). We show that minimizing $D_{-1}(\pi_\theta \| \pi_f)$ yields a closed-form training objective that repels π_θ from the forget distribution; we call this objective **Mode-Avoiding Rényi Unlearning (MARU)**. Anchoring this objective at the reference (pre-unlearning) policy, in the spirit of on-policy methods, ensures that utility preservation follows implicitly from the same gradient geometry underlying continual learning regimes, without requiring an explicit retain loss or retain data.

Our contributions are as follows:

(1) A continual learning-based perspective for LLM unlearning. We cast LLM unlearning as the inverse analogue of on-policy, reverse-KL fine-tuning, showing that the negative-order Rényi divergence can act as the repulsive counterpart to the positive-order regime.

(2) MARU, a mode-avoiding, retain-free unlearning objective. We derive a novel, closed-form unlearning objective by minimizing $D_{-1}(\pi_\theta \| \pi_f)$, that pushes the model off forget completions while its geometry implicitly bounds collateral forgetting, without an explicit retain regularizer.

(3) Empirical validation. Extensive experiments across the TOFU and MUSE benchmarks show that MARU achieves competitive forgetting while delivering up to $31.5\times$ better resistance to membership inference attacks than comparable distillation-based methods.

2 PRELIMINARIES

2.1 MACHINE UNLEARNING FOR LLMS

Let $\mathcal{V}$ denote the token vocabulary and let θ parameterize an LLM. Given a prompt $x \in \mathcal{V}^*$, the model autoregressively generates a response $y \in \mathcal{V}^*$ of length $|y|$. At decoding step $i \in \{1, \ldots, |y|\}$, the model induces a conditional distribution $\pi_\theta(\cdot \mid x, y_{<i})$ over $\mathcal{V}$, where $y_{<i}$ denotes the tokens generated prior to step i. The likelihood of a full response y given x is

$$\pi_\theta(y \mid x) = \prod_{i=1}^{|y|} \pi_\theta(y_i \mid x, y_{<i}). \tag{1}$$

Let $\pi_{\text{ref}} \equiv \pi_{\theta_0}$ denote the pre-unlearning reference policy and π_θ the policy being optimized. We write $\pi_{\text{f}}(y \mid x)$ for the *forget distribution*, the empirical conditional distribution over responses from which the unlearning procedure must remove probability mass.

LLM Unlearning Objective. Given a pre-trained model θ_0 and a *forget set* $\mathcal{D}_{\text{f}} = \{(x, y)\}$ of prompt-response pairs representing knowledge, behaviors, or information to be erased, LLM unlearning seeks parameters θ such that π_θ no longer reliably reproduces y given x for $(x, y) \in \mathcal{D}_{\text{f}}$, nor exhibits the underlying proficiency, while preserving π_θ's behavior on all inputs outside $\mathcal{D}_{\text{f}}$. Following the general formulation adopted across the LLM unlearning literature (Liu et al., 2025), existing methods operationalize unlearning as a two-term objective over $\mathcal{D}_{\text{f}}$ and an auxiliary *retain set* $\mathcal{D}_r$ of general-domain or held-out instances:

$$\min_\theta \quad \underbrace{\mathbb{E}_{\mathcal{D}_{\text{f}}}\big[\mathcal{L}_f(\theta; \mathcal{D}_{\text{f}})\big]}_{\text{Forget loss } \mathcal{L}_f} + \lambda \underbrace{\mathbb{E}_{\mathcal{D}_r}\big[\mathcal{L}_r(\theta; \mathcal{D}_r)\big]}_{\text{Retain loss } \mathcal{L}_r}, \tag{2}$$

where $\mathcal{L}_f$ suppresses proficiency on $\mathcal{D}_{\text{f}}$, $\mathcal{L}_r$ penalizes deviation from π_{ref} on $\mathcal{D}_r$, and λ trades off the two (Yuan et al., 2025). Utility preservation in this formulation is *extrinsic*: it is imposed by $\mathcal{L}_r$ rather than following from any property of $\mathcal{L}_f$ itself, which requires access to retain data and an additional hyperparameter λ.

2.2 f-DIVERGENCE AND RÉNYI DIVERGENCE

A divergence $D(P\|Q)$ is a way of measuring how far a distribution P is from a reference distribution Q. The f-divergences are the standard way of constructing such a measure from a single convex function. Given two distributions P, Q over a common support with P absolutely continuous with respect to Q, an f-divergence is defined as

$$D_f(P \,\|\, Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right], \tag{3}$$

for a convex function f with $f(1) = 0$. Different choices of f recover familiar divergences: $f(t) = t \log t$ gives the KL divergence $\mathrm{KL}(P\|Q)$, and $f(t) = (t-1)^2$ gives the χ^2-divergence.

Rényi divergence. A closely related family, indexed by an order $\alpha \in \mathbb{R} \setminus \{1\}$, is the Rényi divergence

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^\alpha\right], \tag{4}$$

which reduces to the KL divergence in the limit $\alpha \to 1$, i.e., $\lim_{\alpha\to 1} D_\alpha(P\|Q) = \mathrm{KL}(P\|Q)$ (Rényi, 1961). Other notable special cases include the Hellinger distance at $\alpha = 1/2$ and $D_2(P\|Q) = \log(1 + \chi^2(P\|Q))$ at $\alpha = 2$.

3 Related Work

3.1 LLM Unlearning

With the rapid adoption and progress of LLMs, unlearning in these systems has become a critical research area (Geng et al., 2025; Liu et al., 2024; 2025; Yao et al., 2024a). In contrast to earlier work centered on formal guarantees, research on LLM unlearning has largely prioritized empirically grounded methods that scale in practice. The resulting landscape of methods can be broadly grouped into three families. The first builds on gradient-based objectives, such as GA, GD and Weighted Gradient Ascent (WGA), to reverse or negate the training signal associated with the data to be forgotten (Eldan & Russinovich, 2023; Jang et al., 2023). The second draws on preference-optimization frameworks, adapting alignment-style objectives such as Direct Preference Optimization (DPO), NPO, SimNPO to suppress unwanted knowledge (Rafailov et al., 2023; Zhang et al., 2024; Mekala et al., 2025; Fan et al., 2026; Zhang et al., 2025a; Zaradoukas et al., 2026b;a). The third takes a distillation-based approach, methods like Unlearning via Self-Distillation on Adjusted Logits (UnDIAL) or Distilled Unlearning from an Efficient Teacher (DUET) construct a teacher distribution with the undesired information removed and train a student model to reproduce this modified behavior (Dong et al., 2025; Zhong et al., 2026; Lee et al., 2026). Because these techniques are inherently empirical, a substantial body of work has emerged on evaluating LLM unlearning (Maini et al., 2024; Shi et al., 2025; Li et al., 2024; Dorna et al., 2026) and critiquing the limitations of current evaluation protocols (Scholten et al., 2025; Yuan et al., 2025).

3.2 Continual Learning in LLMs

Classical continual learning addresses catastrophic forgetting through regularization toward important parameters (Kirkpatrick et al., 2017), replay of past data (Rebuffi et al., 2017), or architectural isolation of task-specific capacity (Yoon et al., 2018; Rusu et al., 2016). Recent work on LLMs instead traces forgetting to the geometry of the fine-tuning objective: Shenfeld et al. (2026b) show that on-policy RL fine-tuning forgets less than SFT at matched task accuracy, attributing this to an implicit bias of on-policy sampling toward KL-minimal solutions. Building on knowledge distillation (Hinton et al., 2015), a line of self-distillation methods extends this on-policy, reverse-KL benefit to the self-distillation settings (Shenfeld et al., 2026a; Hübotter et al., 2026; Zhao et al., 2026; Ye et al., 2026; Buening et al., 2026). None of this work, however, considers the opposite problem: removing rather than reinforcing a mode. We show that taking advantage of this implicit capability-preservation property, and tweaking it, yields a principled objective for unlearning.

3.3 Divergence Families as Training Objectives

Beyond KL, α-indexed divergence families have been used directly as optimization objectives, principally in approximate Bayesian inference: black-box α-divergence minimization scales α-divergence fitting via stochastic gradient descent (Hernandez-Lobato et al., 2016), adversarial α-divergence minimization extends this to implicit posteriors via a discriminator (Rodriguez-Santana & Hernández-Lobato, 2022), and Rényi variational inference develops corresponding bounds for the logarithmic parameterization (Li & Turner, 2016). Across this line, the order is treated as an interpolation knob between mass-covering and mode-seeking fits and is confined to the attractive regime: Hernandez-Lobato et al. (2016) and Rodriguez-Santana & Hernández-Lobato (2022) search only $\alpha \in (0, 1]$, and while Li & Turner (2016) discuss negative orders, they do not train with them. Novello et al. (2026) unify diffusion-model unlearning under a general f-divergence framework, but remain restricted to the attractive regime and the image domain. We are, to our knowledge, the first to use the negative-order regime as a training objective, and the first to identify it as the natural geometry for unlearning.

4 MODE-AVOIDING UNLEARNING: A CONTINUAL LEARNING PERSPECTIVE

Motivated by the recent line of continual learning literature mentioned in Section 3, we argue that unlearning can be viewed as the inverse analogue of continual learning: rather than adding a mode while preserving the rest of the distribution, we must *remove* a designated forget mode while leaving the rest untouched. We now make this precise by proposing **MARU** (**M**ode-**A**voiding **R**ényi **U**nlearning). We first show that no positive-order Rényi divergence can unlearn (§4.1); derive the negative-order objective that flips attraction into repulsion (§4.2); turn this equivalence into a practical, trainable loss (§4.3); and show that the reference-normalized weighting empirically yields implicit utility retention (§5).

4.1 NO POSITIVE ORDER UNLEARNS

In the continual learning setting the reverse KL objective is responsible for the utility-preserving geometry of on-policy fine-tuning: minimizing $\mathrm{KL}(\pi_\theta \| \pi_t)$ against a fixed target π_t concentrates updates on π_t's high-density regions, which is what leaves prior knowledge intact (Shenfeld et al., 2026b;a). Unlearning cannot reuse this objective directly, however: reverse KL is still *attractive*, minimized when π_θ matches its target, and no relabeling of the target turns an attractive objective into a repulsive one. What we need is not a different target but a different kind of objective, so we look for one within a more general family that reverse KL is already known to belong to.

The KL divergence is the $\alpha \to 1$ member of exactly such a family, the Rényi divergence, and this is our first lever: rather than fixing the divergence and asking what target to move toward, we fix π_θ against the forget distribution and ask whether *some order* α makes the divergence itself repulsive. Recall from §2.2 that $D_\alpha(\pi_\theta \| \pi_\mathrm{f}) \to \mathrm{KL}(\pi_\theta \| \pi_\mathrm{f})$ as $\alpha \to 1$, and that for $\alpha > 0$, increasing α moves the matching behavior of D_α from mode-covering toward mode-seeking. What this spectrum obscures is that *every* positive order is minimized by the same distribution.

Proposition 1 (No positive order unlearns). *For every $\alpha > 0$ and all $\pi_\theta, \pi_\mathrm{f}$ with $\pi_\theta \ll \pi_\mathrm{f}$,*

$$D_\alpha(\pi_\theta \| \pi_\mathrm{f}) \geq 0, \tag{5}$$

with equality if and only if $\pi_\theta = \pi_\mathrm{f}$ almost everywhere (Van Erven & Harremos, 2014).

Consequently, $\min_\theta D_\alpha(\pi_\theta \| \pi_\mathrm{f})$ is an *attractive* objective for every positive order: no choice of $\alpha > 0$ can produce an objective whose minimum pushes π_θ *away* from π_f. Varying α within this regime changes how mass is matched, mode-covering as $\alpha \to 0^+$, increasingly mode-seeking as α grows (Li & Turner, 2016), but never whether it is matched. So staying inside the positive-order regime is not enough on its own: repulsion requires leaving it, which is the move we make next.

4.2 FROM ATTRACTION TO REPULSION: THE NEGATIVE-ORDER REGIME

Leaving the positive regime means asking what $D_\alpha(\pi_\theta \| \pi_\mathrm{f})$ becomes for $\alpha < 0$, and here two things are not yet clear: whether the functional stays a sensible object to optimize at all outside $(0, \infty)$, and, if so, whether descending it actually pushes π_θ away from π_f rather than simply becoming undefined or degenerate. Both questions are answered by relating negative orders back to the positive-order divergences already known to be attractive, through a skew symmetry of the Rényi family.

Proposition 2 (Skew symmetry). *For all admissible $\alpha \neq 0, 1$,*

$$D_\alpha(\pi_\theta \| \pi_\mathrm{f}) = \frac{\alpha}{1-\alpha} D_{1-\alpha}(\pi_\mathrm{f} \| \pi_\theta). \tag{6}$$

The full proof is deferred to Appendix A.1.

Corollary 1 (The sign flip). *For $\alpha < 0$ we have $1 - \alpha > 1$, so $D_{1-\alpha}(\pi_\mathrm{f} \| \pi_\theta) \geq 0$ is a genuine Rényi divergence while $\frac{\alpha}{1-\alpha} < 0$; hence $D_\alpha(\pi_\theta \| \pi_\mathrm{f}) \leq 0$. Since this prefactor is a fixed negative constant for a given α, minimizing the left side of equation 6 over θ is equivalent to maximizing the right side:*

$$\min_\theta D_\alpha(\pi_\theta \| \pi_\mathrm{f}) \iff \max_\theta D_{1-\alpha}(\pi_\mathrm{f} \| \pi_\theta). \tag{7}$$

Maximizing $D_{1-\alpha}(\pi_\mathrm{f} \| \pi_\theta)$ with order $1 - \alpha > 1$ is an inclusive divergence of the forget distribution from the model: maximizing it forces π_θ to place little mass where π_f places mass, precisely the mode-avoiding behavior we want.

Corollary 2 (The $\alpha = -1$ specialization)**.** *Setting $\alpha = -1$ in equation 7 gives $1 - \alpha = 2$, so, using $D_2(P\|Q) = \log(1 + \chi^2(P\|Q))$ from §2.2,*

$$\min_\theta D_{-1}(\pi_\theta\|\pi_\mathrm{f}) \iff \max_\theta D_2(\pi_\mathrm{f}\|\pi_\theta) = \max_\theta \log\big(1 + \chi^2(\pi_\mathrm{f}\|\pi_\theta)\big). \tag{8}$$

Minimizing the negative-order Rényi divergence [1] is thus exactly equivalent to maximizing a χ^2-type divergence of the forget distribution from the model, a well-understood mode-avoiding objective, obtained without introducing any new machinery beyond the Rényi family already used to describe on-policy fine-tuning.

4.3 The MARU Objective

By definition (§2.2), the MARU objective is the order-α Rényi divergence of the model from the forget distribution itself,[2]

$$\mathcal{L}(\theta) \;=\; D_\alpha(\pi_\theta\|\pi_\mathrm{f}), \qquad \alpha < 0. \tag{9}$$

By Corollary 1, minimizing this objective is equivalent to maximizing the proper, non-negative order-$(1-\alpha)$ divergence $D_{1-\alpha}(\pi_\mathrm{f}\|\pi_\theta)$, and is therefore mode-avoiding for every $\alpha < 0$. What remains is to turn Equation 9 into a quantity we can compute from the forget set $\mathcal{D}_\mathrm{f}$ alone.

The forget set $\mathcal{D}_\mathrm{f} = \{(x_i, y_i)\}_{i=1}^N$ defines an empirical forget distribution

$$Q_f \;=\; \frac{1}{N}\sum_{i=1}^N \delta_{(x_i,y_i)}. \tag{10}$$

That is, Q_f is the empirical estimate of the forget distribution π_f from $\mathcal{D}_\mathrm{f}$. We keep Q_f distinct from the pre-unlearning reference policy π_ref: Q_f specifies which prompt–completion pairs are designated for forgetting, while π_ref provides a fixed reference likelihood against which changes in the model are measured.

To turn Equation 9 into a quantity computable from $\mathcal{D}_\mathrm{f}$ alone, we make two substitutions. First, since π_f has no accessible closed form, we instantiate its density using the reference model itself, $\pi_\mathrm{f}(y \mid x) := \pi_\mathrm{ref}(y \mid x)$ for x in the forget prompt set, a natural choice since π_ref is precisely the pre-unlearning model that assigned these completions high likelihood. Second, we replace the resulting expectation under π_f with the empirical forget distribution Q_f. Both steps are approximations of Equation 9, not identities; the propositions of §4.2 characterize $D_\alpha(\pi_\theta\|\pi_\mathrm{f})$ exactly, and the loss below is a plug-in estimator of it under $\pi_\mathrm{f} \approx \pi_\mathrm{ref}$. Combining both substitutions,

$$\mathcal{L}(\theta) \;=\; \frac{1}{\alpha - 1}\log \mathbb{E}_{(x,y)\sim Q_f}\left[\left(\frac{\pi_\theta(y \mid x)}{\pi_\mathrm{ref}(y \mid x)}\right)^\alpha\right], \qquad \alpha < 0. \tag{11}$$

Since Q_f is the empirical distribution over $\mathcal{D}_\mathrm{f}$, this expectation is exact rather than estimated, and evaluates to

$$\mathcal{L}(\theta) = \frac{1}{\alpha - 1}\log\left(\frac{1}{|\mathcal{D}_\mathrm{f}|}\sum_{(x_i,y_i)\in\mathcal{D}_\mathrm{f}} e^{\alpha\rho_i(\theta)}\right), \qquad \alpha < 0. \tag{12}$$

where $\rho_i(\theta) = \log\frac{\pi_\theta(y_i|x_i)}{\pi_\mathrm{ref}(y_i|x_i)}$ is the reference-normalized log-likelihood ratio of forget example i. In principle, Equation 12 is unbounded below: as training pushes $\pi_\theta(y_i \mid x_i) \to 0$ for some forget example, precisely the direction the objective rewards, $\rho_i(\theta) \to -\infty$, and since $\alpha < 0$ this gives $\alpha\rho_i(\theta) \to +\infty$, $e^{\alpha\rho_i(\theta)} \to +\infty$, and, because $1/(\alpha - 1) < 0$, $\mathcal{L}(\theta) \to -\infty$. A single example driven to sufficiently low likelihood under π_θ could in principle dominate the sum and send the loss to $-\infty$. As a safeguard, we tried clipping each log-ratio from below before exponentiating

[1] $D_\alpha(\pi_\theta\|\pi_\mathrm{f})$ is non-positive for $\alpha < 0$ and thus not a divergence in the information-theoretic sense; it is a signed repulsion functional, restated rigorously via Proposition 2.

[2] It is essential that this is the *logarithmic* Rényi form rather than the non-logarithmic Tsallis/Amari α-divergence commonly listed in f-divergence references: the two agree in sign only for $\alpha \in (0, 1]$ and disagree for $\alpha < 0$, so substituting the latter at $\alpha = -1$ would train π_θ *toward* π_f. See Appendix A.2 for details.

(Appendix A.3), but found this to be ineffective, never improving on the best forgetting–utility trade-off of the unclipped run (Figure 5). We therefore train directly on Equation 12, without clipping, for all reported results.

Notably, unlike the standard formulation in Equation 2, Equation 12 contains no retain term $\mathcal{L}_r$ and no trade-off weight λ. Utility preservation instead falls out of the same geometry that causes on-policy RL to forget less than SFT, without ever computing a retain loss or requiring retain data.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmarks We evaluate MARU on two standard LLM unlearning benchmarks integrated within the OPENUNLEARNING framework (Dorna et al., 2026) to ensure a consistent evaluation protocol. TOFU (Maini et al., 2024) consists of 200 fictitious author profiles with synthetic QA pairs, split into Forget 1%/5%/10% subsets designating $\mathcal{D}_{\mathrm{f}}$ against the remaining retain split $\mathcal{D}_r$. We report results on Llama-3.2-1B-Instruct in the main text, with 3B and 7B results deferred to Appendix B. We further evaluate on MUSE (Shi et al., 2025), which targets long-form, verbatim memorization on real-world NEWS and BOOKS corpora with disjoint forget/retain splits, using Llama-2-7b-hf, to test whether MARU's mode-avoiding geometry generalizes beyond TOFU's setting.

Baseline Methods We compare MARU against baselines spanning the major unlearning paradigms, all run within OPENUNLEARNING under matched compute: gradient-based methods GA, GD, and WGA; preference-optimization methods NPO, SimNPO, and DPO; the representation-based method RMU; and distillation-based methods UNDIAL, DUET, Off-Policy Distillation (OPD), and Self-Distillation Finetuning (SDFT) (see Appendix C.1 for details). Alongside these baselines, we report two reference points for every setting: the TARGET model, the pre-unlearning checkpoint from which each method in Table 1 starts, serving as a lower bound; and the RETRAINED model, which approximates a model trained without the forget set and serves as an oracle upper bound that unlearning methods aim to approximate without incurring the cost of retraining. Method-specific hyperparameters follow the original works; further details are given in Appendix C.2.

Metrics Following OPENUNLEARNING, we report four metrics per split. The Memorization (MEM) metric is the Harmonic Mean (HM) of four TOFU unlearning metrics, measuring suppression of forget-set knowledge. The Privacy (PRIV) metric is the HM of four membership-inference attack scores (LOSS, Zlib, Min-k, and Min-k++). The Model Utility (UTIL) metric aggregates ten TOFU metrics spanning the retain, real-author, and world-fact splits. Finally, AGG, the HM of MEM and UTIL, summarizes the forgetting–utility trade-off; we use it to select the best checkpoint per method, consistent with Tables 1–4. For MUSE, we report VerbMem (VM) and KnowMem (KM), both for the forget/retain sets, as defined in Shi et al. (2025). Full definitions of all metrics appear in Appendix D.2.

5.2 RESULTS

MARU achieves SOTA performance on the hardest forget split. Table 1 reports TOFU unlearning results for Llama-3.2-1B-Instruct at forget ratios of 1%, 5%, and 10%. Every method improves markedly over the TARGET model, whose MEM scores of 0.064–0.088 confirm near-complete retention of the forget set, but none matches the RETRAINED oracle (AGG 0.616–0.643); together these define the lower and upper bounds of unlearning performance on this benchmark. The results expose a clear trade-off axis: gradient- and preference-based methods (GA, GD, NPO) attain the strongest forgetting but forfeit Utility once the forget set grows beyond 1%, whereas WGA, SimNPO, and the self-distillation family (OPD, UNDIAL, DUET, SDFT) preserve Utility close to the TARGET model at the cost of substantially weaker forgetting (Mem ≤ 0.40). MARU traces a different path along this frontier: at Forget 1% it achieves the second best Utility in the table (0.726), and at Forget 10% it is the strongest method overall, leading on AGG (0.546), MEM (0.560, within 0.024 of the oracle), and PRIV (0.481). This comes at a cost in MARU's Utility at Forget 10%, where it drops to 0.533, meaning that its advantage at the largest forget ratio is best read as better trade-off of general utility for forgetting performance.

Table 1: TOFU unlearning results for Llama-3.2-1B-Instruct. Best checkpoint per method chosen by HM(Mem, Util). All metrics are higher-is-better. **Bold** = best method per column, underlined = second best.

	Forget 1%				Forget 5%				Forget 10%			
Method	Agg	Mem	Priv	Util	Agg	Mem	Priv	Util	Agg	Mem	Priv	Util
Target	0.118	0.064	0.004	0.718	0.153	0.086	0.004	0.706	0.157	0.088	0.004	0.708
Retrained	0.616	0.546	—	0.708	0.641	0.577	—	0.722	0.643	0.584	—	0.715
GA	**0.536**	0.429	**0.533**	0.713	**0.546**	**0.507**	0.316	0.590	0.524	0.461	0.171	0.607
GD	0.509	0.414	0.447	0.661	0.520	0.462	**0.367**	0.595	0.519	0.468	0.373	0.581
WGA	0.514	0.400	0.067	0.716	0.473	0.354	0.045	**0.711**	0.464	0.344	0.032	**0.713**
NPO	0.530	0.423	0.315	0.707	0.545	0.485	0.232	0.622	0.517	0.445	0.137	0.617
SimNPO	0.246	0.149	0.007	0.721	0.219	0.129	0.006	**0.711**	0.227	0.135	0.005	0.708
DPO	0.406	0.284	0.021	0.711	0.418	0.305	0.033	0.665	0.443	0.330	0.036	0.676
RMU	0.533	**0.440**	0.153	0.675	0.499	0.414	0.080	0.628	0.526	0.440	0.150	0.655
OPD	0.363	0.242	0.029	0.724	0.353	0.235	0.018	0.705	0.357	0.239	0.018	0.704
UNDIAL	0.477	0.355	0.024	**0.727**	0.443	0.326	0.011	0.691	0.446	0.330	0.006	0.688
DUET	0.236	0.141	0.007	0.715	0.205	0.120	0.005	0.702	0.222	0.131	0.006	0.704
SDFT	0.397	0.275	0.032	0.714	0.436	0.332	0.029	0.635	0.440	0.332	0.031	0.653
MARU	0.455	0.331	0.299	0.726	0.471	0.362	0.121	0.676	**0.546**	**0.560**	**0.481**	0.533

MARU resists privacy leakage where distillation baselines fail. Beyond the MEM metric, Tables 1–4 reveal a second, orthogonal effect: on the PRIV metric, MARU not only outperforms the distillation baselines, most similar to it in motivation (UNDIAL, DUET, OPD, SDFT), it does so by a wide and consistent margin. Figure 2 isolates this comparison across all three model scales and all three forget ratios. While UNDIAL, DUET, OPD, and SDFT all cluster close to zero, indicating that a model unlearned with these methods remains highly identifiable as having seen the forget set, MARU achieves a PRIV score $5.0\times$ to $31.5\times$ higher than the mean of these baselines in every one of the nine settings. We attribute this to a structural difference rather than a tuning artifact: UNDIAL, DUET, OPD, and SDFT all construct a fixed teacher distribution and pull π_θ toward it, so any residual similarity between the teacher and the true forget distribution is directly inherited by the student. MARU instead optimizes a signed repulsion functional (§4.2) that actively pushes π_θ away from the empirical forget distribution, which is reflected in weaker membership-inference identifiability without ever training against an explicit privacy objective.

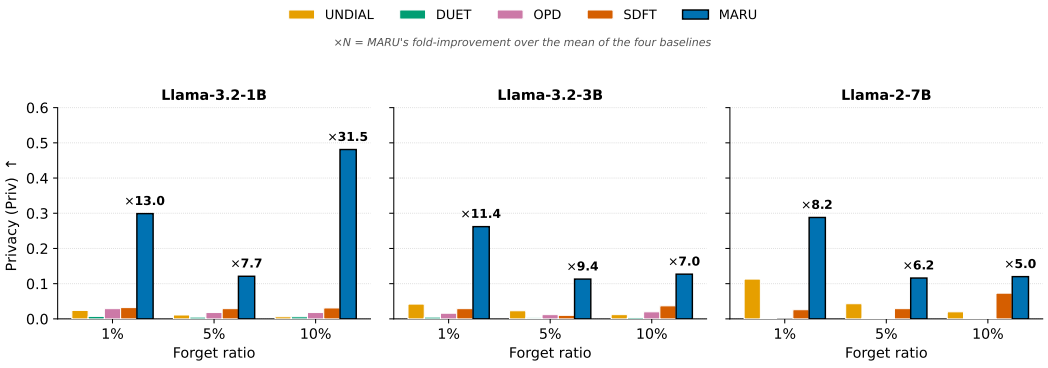


Figure 2: Privacy scores across model scales and forget ratios. MARU improves $5.0\times$–$31.5\times$ over distillation baselines (UNDIAL, DUET, OPD, SDFT) on average, showing that MARU's mode-avoiding geometry protects privacy as a byproduct of forgetting rather than an explicit constraint.

MARU's forgetting geometry transfers less cleanly to long-form, verbatim memorization. Table 2 reports MUSE unlearning results for Llama-2-7b-hf on NEWS and BOOKS. Unlike the TOFU setting, no single method dominates across all metrics, and MARU's performance here sits at the middle rather than being decisive: on NEWS, it reduces VerbMem$_F$ from 57.1 (Target) to 28.9,

Table 2: MUSE results for Llama-2-7b-hf on News and Books. VM = VerbMem, KM = KnowMem; subscripts F/R denote forget/retain sets. Lower is better for VM_F and KM_F, higher for KM_R. Values are reported in %. **Bold** = best method per column, underline = second best.

	News			Books		
Method	VM_F	KM_F	KM_R	VM_F	KM_F	KM_R
Target	57.1	65.4	54.4	99.7	46.6	68.1
Retrain	20.4	32.1	56.3	14.4	30.3	69.6
GA	23.1	57.6	46.2	59.5	31.7	54.0
GD	**22.7**	53.0	43.8	60.0	35.7	55.6
SimNPO	53.0	59.2	49.7	27.4	**27.2**	59.5
RMU	38.0	52.6	49.2	**20.2**	33.3	57.7
UNDIAL	24.4	**25.8**	38.3	59.5	40.3	62.2
DUET	56.6	63.6	**55.9**	99.7	47.0	**69.4**
MARU	28.9	58.6	49.4	68.6	35.0	55.9

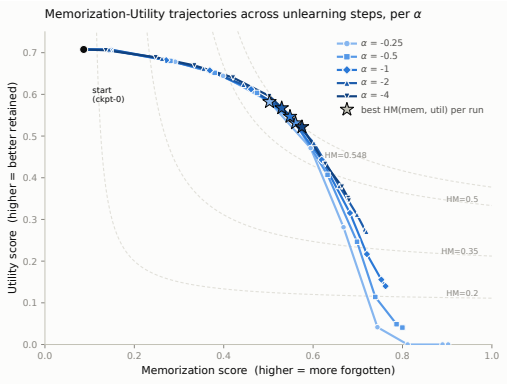


Figure 3: Memorization–utility trajectories traced by MARU, for different values of α. Dashed gray curves are iso-contours of $\text{HM}(\text{memorization}, \text{utility})$ on TOFU 10% split.

comparable to GA and GD, while its KnowMem_F (58.6) stays much closer to Target (65.4) than to the strongest baseline (UNDIAL, 25.8), suggesting MARU suppresses verbatim recall more effectively than paraphrased knowledge; its KnowMem_R (49.4) sits in the middle, essentially tied with SimNPO (49.7) and RMU (49.2), reflecting retained utility without a clear edge. BOOKS follows a similar pattern of being competitive but not state-of-the-art. Overall, MARU is never the best or second-best method on any individual MUSE column, in contrast to its TOFU results (Table 1), suggesting that the mode-avoiding, retain-free geometry underlying MARU transfers less cleanly to MUSE's long-form, verbatim memorization setting than to TOFU's QA-style forget targets.

The forget–utility trade-off is optimal near $\alpha = -1$. Figure 3 plots each run's trajectory through memorization–utility space as unlearning proceeds. All five runs (sweeping $\alpha \in \{-0.25, -0.5, -1, -2, -4\}$) start from the same target-model checkpoint and initially move along an almost identical curve, indicating that α has little effect on the shape of the early trade-off and mainly rescales the number of steps needed to reach a given level of forgetting. The trajectories then separate sharply: $\alpha = -0.25$, the setting closest to zero, breaks off first, over-forgetting as memorization climbs toward 0.90 while utility collapses to near zero before settling anywhere useful. More negative values of α instead trace flatter, more gradual paths that graze the outer (top-right) frontier for longer. Scoring each run by the harmonic mean of the two objectives, the achievable trade-off is essentially flat across $\alpha \in [-4, -0.5]$, with $\alpha = -1$ and $\alpha = -2$ tied for the global optimum. We therefore recommend operating at $\alpha \in [-2, -1]$, evaluated around the midpoint of training, as the point that forgets the most while sacrificing the least utility.

6 CONCLUSION

In this work, we presented MARU, a mode-avoiding method for LLM unlearning that reframes forgetting as the inverse analogue of on-policy continual learning and leverages the negative-order Rényi divergence to turn an attractive objective into a repulsive one. On TOFU, MARU leads all baselines at Forget 10% at 1B scale and resists membership-inference attacks up to 31.5× better than comparable distillation-based methods, without optimizing a privacy objective, while these advantages narrow at larger scales and diminish when carried over to MUSE. Overall, our findings highlight the promise of treating unlearning as a question of divergence geometry rather than heuristic gradient design, offering a principled and retain-free direction for controllable LLM unlearning, and we believe this paradigm opens new research avenues in the intersection of unlearning and continual learning for LLMs.

AI USE STATEMENT

We used generative AI tools (LLMs) at several stages of this work and disclose that use below:

Tasks with required disclosure. We used generative AI tools to help develop the conceptual framing of unlearning as the inverse analogue of on-policy continual learning (§4); to refine the statements of Proposition 1 and Proposition 2 and to assist in drafting the proof in Appendix A.1; to implement the MARU objective (Eq. 12) within the OPENUNLEARNING framework. Every mathematical claim, derivation, and proof step was independently re-derived and verified by at least two authors, and every quantitative statement in the text was traced back to logged experimental output rather than to model generated summaries. We did not use generative AI tools to generate synthetic data sets, or to design the experimental methodology or evaluation protocol, which follow the OPENUNLEARNING benchmark suite. Assistance with translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work.

Tasks with recommended disclosure. We additionally used generative AI tools to edit author-written drafts, like fixing awkward phrasing, removing redundancy, correcting grammar, and improving paragraph flow, and for general brainstorming during the development of the method. We did not use generative AI tools to produce Figures 1–3, or to identify or summarize related work: all cited references were retrieved and read in their original form by the authors, and no citation appears in this paper solely on the basis of a model generated suggestion. All AI-assisted code was reviewed and tested by the authors.

We have reviewed all AI-assisted work and take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our research focuses on improving machine unlearning to support privacy protection, regulatory compliance, and safer deployment of LLMs. All experiments rely exclusively on publicly available benchmarks, TOFU and MUSE, whose forget sets consist of synthetic profiles of fictitious authors and of publicly available news and book corpora distributed for unlearning research. While machine unlearning can, in principle, be misused (e.g., to suppress or censor inconvenient rather than sensitive content), we emphasize its responsible application as a tool for compliance and user protection. Our method does not introduce new capabilities that expand the risk profile of existing LLMs; like all approximate unlearning methods, however, it offers no formal guarantee of removal, so its empirical scores should not be read as certification of compliance with a deletion request.

REPRODUCIBILITY STATEMENT

We describe the steps taken to ensure our results are reproducible. The full proofs are given in the Appendix A. All experiments are run within the OPENUNLEARNING framework (Dorna et al., 2026) on the publicly available TOFU (Maini et al., 2024) and MUSE (Shi et al., 2025) benchmarks, using their standard forget/retain splits as described in Appendix D.1; we inherit preprocessing, splits, and evaluation code from OPENUNLEARNING to keep MARU and all baselines under a matched evaluation protocol. Full definitions of every reported metric (Mem, Priv, Util, Agg for TOFU; VerbMem and KnowMem for MUSE) are given in Appendix D.2. Baseline formulations (GA, GD, WGA, NPO, SimNPO, DPO, RMU, OPD, UNDIAL, DUET, SDFT) are specified in full in Appendix C.1, and method-specific hyperparameters are reported in Appendix C.2. Compute resources and measured wall-clock training times per method are reported in Appendix C.2 (Table 5), along with the total compute budget across all reported runs. We will release our code implementing MARU within OPENUNLEARNING, along with configuration files for all experiments in the main text and appendix, upon publication; in keeping with the double-blind review process, an anonymous repository link is included instead: `https://anonymous.4open.science/r/maru-B2C4`.

REFERENCES

Thomas Kleine Buening, Jonas Hübotter, Barna Pásztor, Idan Shenfeld, Giorgia Ramponi, and Andreas Krause. Aligning language models from user interactions. In *Continual Adaptation at Scale:*

Towards Sustainable AI Workshop, 2026. URL https://openreview.net/forum?id=HL3iloEczS.

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)*, pp. 2633–2650, 2021.

Howard Chen, Noam Razin, Karthik R Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ODTM64azGa.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8827–8840, 2025.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *Advances in Neural Information Processing Systems*, 38, 2026.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

European Union. Regulation (eu) 2016/679 of the european parliament and of the council. *Official Journal of the European Union*, 2016.

European Union. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *Official Journal of the European Union*, 2023.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *Advances in Neural Information Processing Systems*, 38:1540–1567, 2026.

Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. A comprehensive survey of machine unlearning techniques for large language models. *CoRR*, 2025.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *International conference on machine learning*, pp. 1511–1520. PMLR, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QkfkxyRizZ.

Hugging Face. Sdft trainer. https://huggingface.co/docs/trl/sdft_trainer, 2025. TRL documentation.

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 *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Madhur Jindal. Gibberish detector: High-accuracy text classification model, 2021. URL https://huggingface.co/madhurjindal/autonlp-Gibberish-Detector-492513457.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Bruce W Lee, Addie Foote, Alex Infanger, Leni Shor, Harish Kamath, Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, and Alexander Turner. Distillation robustifies unlearning. *Advances in Neural Information Processing Systems*, 38:88615–88649, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 28525–28550, 2024.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.

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*, pp. 1–14, 2025.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Machine unlearning in generative ai: A survey. *CoRR*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. In *First Conference on Language Modeling*, 2024.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3732–3752, 2025.

Nicola Novello, Federico Fontana, Luigi Cinque, Deniz Gunduz, and Andrea M Tonello. A unified framework for diffusion model unlearning with f-divergence. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=6kyKD3bVi6.

Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68, 2018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5533–5542. IEEE, 2017.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Simon Rodriguez-Santana and Daniel Hernández-Lobato. Adversarial α-divergence minimization for bayesian approximate inference. *Neurocomputing*, 471:260–274, 2022.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Yan Scholten, Stephan Günnemann, and Leo Schwinn. A probabilistic perspective on unlearning and alignment for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations*, volume 2026, pp. 59839–59864, 2026b.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024.

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. In *The Thirteenth International Conference on Learning Representations*, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European symposium on security and privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=u3vEuRr08MT.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE transactions on information theory*, 60(7):3797–3820, 2014.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=huo8MqVH6t.

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

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=8Dy42ThoNe.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sk7KsfW0-.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *International Conference on Learning Representations*, volume 2025, pp. 49483–49508, 2025.

Efstratios Zaradoukas, Davide Gabrielli, Bardh Prenkaj, and Gjergji Kasneci. Beyond binary rewards: A comparative study of reward design for reinforcement unlearning. *arXiv preprint arXiv:2607.27968*, 2026a.

Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Reinforcement unlearning via group relative policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=BjWwqPE7mk.

Chenlong Zhang, Zhuoran Jin, Hongbang Yuan, Jiaheng Wei, Tong Zhou, Kang Liu, Jun Zhao, and Yubo Chen. Rule: Reinforcement unlearning achieves forget-retain pareto optimality. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. Min-k%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, volume 2025, pp. 64845–64862, 2025b.

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

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

Yisheng Zhong, Zhengbang Yang, and Zhuangdi Zhu. DUET: Distilled LLM unlearning from an efficiently contextualized teacher. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Xa6QRrXrKX.

A Proofs and Additional Analysis

A.1 Proof of Proposition 2

By Equation 4, for any admissible $\alpha \neq 0, 1$ and distributions Q, P with $Q \ll P$,

$$D_\alpha(Q\|P) = \frac{1}{\alpha - 1} \log \int Q(x)^\alpha P(x)^{1-\alpha}\, dx. \tag{13}$$

Apply this definition to $D_{1-\alpha}(P\|Q)$, i.e., substitute order $1 - \alpha$ and swap the roles of the two arguments:

$$D_{1-\alpha}(P\|Q) = \frac{1}{(1-\alpha) - 1} \log \int P(x)^{1-\alpha} Q(x)^{1-(1-\alpha)}\, dx = \frac{1}{-\alpha} \log \int P(x)^{1-\alpha} Q(x)^\alpha\, dx. \tag{14}$$

Multiplying both sides by $\alpha/(1-\alpha)$,

$$\frac{\alpha}{1-\alpha} D_{1-\alpha}(P\|Q) = \frac{\alpha}{1-\alpha} \cdot \frac{1}{-\alpha} \log \int P(x)^{1-\alpha} Q(x)^\alpha\, dx = \frac{1}{\alpha - 1} \log \int Q(x)^\alpha P(x)^{1-\alpha}\, dx, \tag{15}$$

where the sign flip $\frac{\alpha}{1-\alpha} \cdot \frac{1}{-\alpha} = \frac{1}{\alpha-1}$ is immediate, and the integral on the right is identical to the one in the definition of $D_\alpha(Q\|P)$ (multiplication of the two integrands commutes). This is exactly $D_\alpha(Q\|P)$, proving Eq. 6 with $Q = \pi_\theta$, $P = \pi_\mathrm{f}$.

For the consequence: since $D_{1-\alpha}(\pi_\mathrm{f}\|\pi_\theta) \geq 0$ whenever $1-\alpha \in (0,1) \cup (1,\infty)$ (any genuine Rényi divergence is non-negative by Gibbs' inequality applied to the underlying f-divergence representation), and $\alpha < 0 \Rightarrow \frac{\alpha}{1-\alpha} < 0$ (numerator negative, denominator $1 - \alpha > 1$ positive), the product in Eq. 6 is ≤ 0, so $D_\alpha(\pi_\theta\|\pi_\mathrm{f}) \leq 0$ for all $\alpha < 0$. Because $\frac{\alpha}{1-\alpha} < 0$ is a fixed negative constant for a given α, minimizing $D_\alpha(\pi_\theta\|\pi_\mathrm{f})$ over θ is equivalent to *maximizing* $D_{1-\alpha}(\pi_\mathrm{f}\|\pi_\theta)$ over θ, which is Eq. 7.

A.2 The Rényi vs. Tsallis/Amari α-Family

The object typically listed as "the α-divergence" in standard f-divergence references is the *non-logarithmic* generator

$$f_\alpha(t) = \frac{t^\alpha - \alpha t - (1-\alpha)}{\alpha(\alpha - 1)}, \qquad D_{f_\alpha}(P\|Q) = \frac{1}{\alpha(\alpha-1)} \left(\int p(x)^\alpha q(x)^{1-\alpha}\, dx - 1 \right), \tag{16}$$

which is convex for every α and therefore a genuine, non-negative divergence even for $\alpha < 0$. It obeys the prefactor-free skew symmetry $D_{f_\alpha}(P\|Q) = D_{f_{1-\alpha}}(Q\|P)$, note the absence of the

$\alpha/(1-\alpha)$ prefactor that appears in Proposition 2. The Rényi divergence is the monotone log-transform of this same integral,

$$D_\alpha^R(P\|Q) = \frac{1}{\alpha-1}\log\big(1+\alpha(\alpha-1)\,D_{f_\alpha}(P\|Q)\big). \tag{17}$$

The two families agree at the KL endpoints ($\alpha \to 1$ forward, $\alpha \to 0$ reverse) but *diverge in sign behavior* for $\alpha < 0$. Concretely, at $\alpha = -1$,

$$D_{f_{-1}}(\pi_\theta\|\pi_\mathrm{f}) = \tfrac{1}{2}\,\chi^2(\pi_\mathrm{f}\|\pi_\theta) \;\geq 0, \qquad \text{minimizing} \;\Rightarrow\; \pi_\theta \to \pi_\mathrm{f} \text{ (attraction)}, \tag{18}$$

$$D_{-1}^R(\pi_\theta\|\pi_\mathrm{f}) = -\tfrac{1}{2}\log\big(1+\chi^2(\pi_\mathrm{f}\|\pi_\theta)\big) \;\leq 0, \qquad \text{minimizing} \;\Rightarrow\; \max\chi^2 \text{ (repulsion)}. \tag{19}$$

Both are built from the same underlying $\chi^2(\pi_\mathrm{f}\|\pi_\theta)$: the non-logarithmic form is minimized by it, the Rényi (log) form is maximized by it. **The MARU objective must therefore use the Rényi (log) form**. The distinction has not previously required comment because existing α-divergence objectives operate at orders where it is immaterial: Hernandez-Lobato et al. (2016) and Rodriguez-Santana & Hernández-Lobato (2022) both restrict the search to $\alpha \in (0, 1]$, a range over which the two parameterizations are monotonically related *with the same sign* and therefore induce identical optima. It becomes decisive only at $\alpha < 0$, which is exactly where MARU operates; naively importing "the α-divergence" from a standard f-divergence table and setting $\alpha = -1$ trains π_θ *toward* π_f, the exact opposite of forgetting, despite the two objectives sharing a name and a value of α.

A.3 A Clipped Variant of the MARU Objective

As noted in §4.3, Equation 12 is unbounded below in principle. As a safeguard against this, we also implemented a clipped variant: each log-ratio is clipped from below at a threshold $c < 0$ before exponentiating,

$$\boxed{\mathcal{L}_{\alpha,c}(\theta) = \frac{1}{\alpha-1}\log\left(\frac{1}{N}\sum_{i=1}^{N} e^{\alpha\tilde{\rho}_i(\theta)}\right), \qquad \alpha < 0, \qquad \tilde{\rho}_i(\theta) = \max\{\rho_i(\theta),\, c\}.} \tag{20}$$

Because $\tilde{\rho}_i(\theta) \geq c$ for every i and $\alpha < 0$, each term satisfies $e^{\alpha\tilde{\rho}_i(\theta)} \leq e^{\alpha c}$, so $\mathcal{L}_{\alpha,c}$ is bounded below by $\frac{\alpha}{\alpha-1}\,c$. Operationally, once an example's likelihood ratio has fallen below e^c, it is treated as forgotten and can no longer pull the loss further down.

B Additional Experiments

We repeat the Llama-3.2-1B-Instruct evaluation of §5.2 at two larger scales, Llama-3.2-3B-Instruct (Table 3) and Llama-2-7b-chat-hf (Table 4), to test whether MARU's mode-avoiding geometry is specific to the smallest model or persists as capacity grows.

Larger models sharpen the forgetting–utility trade-off, and MARU's Agg/Mem lead disappears. At both 3B and 7B, the TARGET checkpoint's baseline Utility sits noticeably higher than at 1B (0.72–0.76 vs. 0.71), and several baselines close the gap to MARU on AGG and MEM at every split. UNDIAL's behavior now diverges by scale rather than uniformly strengthening with it: at 7B it is best or second-best on AGG at all three forget ratios, and on MEM at Forget 1% and Forget 5%, only dropping to third on MEM at Forget 10%. At 3B, WGA leads AGG and MEM only at Forget 1%; from Forget 5% onward NPO overtakes it on both metrics. This differs from the 1B setting (Table 1), where GA remains the strongest baseline on AGG and MEM through Forget 5%, with MARU only overtaking it at Forget 10%. MARU is not the best performing method at either larger scale; it remains broadly competitive with the gradient- and preference-based methods (GA, GD, NPO, WGA) rather than falling toward the weakest baselines (SimNPO, DUET, OPD), and on Llama-2-7b-chat-hf it is second-best on UTIL at Forget 1% (0.738) and second-best on PRIV at Forget 10% (0.120).

Privacy resistance persists, but its margin shrinks with scale. Computing the same ratio against the mean PRIV score of the distillation baselines (UNDIAL, DUET, OPD, SDFT) as in §5.2 (Figure 2), MARU retains a consistent advantage at every scale and forget ratio we test: roughly $7.0\times$–$11.4\times$ at 3B and $5.0\times$–$8.2\times$ at 7B. The margin is largest at 1B (peaking at $31.5\times$ on Forget 10%)

Table 3: TOFU unlearning results for Llama-3.2-3B-Instruct. Best checkpoint per method chosen by HM(Mem, Util). **Bold** = best method per column, underlined = second best.

	Forget 1%				Forget 5%				Forget 10%			
Method	Agg	Mem	Priv	Util	Agg	Mem	Priv	Util	Agg	Mem	Priv	Util
Target	0.024	0.012	0.002	0.763	0.057	0.030	0.002	0.749	0.062	0.032	0.002	0.751
Retrained	0.635	0.542	—	0.766	0.648	0.569	—	0.752	0.650	0.574	—	0.749
GA	0.511	0.386	**0.468**	0.756	0.513	0.403	0.146	0.709	0.525	0.432	0.134	0.669
GD	0.489	0.365	0.453	0.738	0.487	0.382	**0.234**	0.673	0.502	0.411	**0.295**	0.645
WGA	**0.527**	**0.408**	0.187	0.743	0.444	0.316	0.038	0.747	0.449	0.320	0.041	0.750
NPO	0.513	0.386	0.282	0.765	**0.516**	**0.406**	0.103	0.707	**0.557**	**0.486**	0.170	0.651
SimNPO	0.226	0.133	0.009	0.750	0.149	0.083	0.004	0.744	0.161	0.090	0.005	0.746
DPO	0.324	0.206	0.013	**0.770**	0.368	0.246	0.017	0.730	0.374	0.252	0.023	0.721
RMU	0.312	0.195	0.010	0.767	0.364	0.241	0.012	0.747	0.434	0.306	0.028	0.745
OPD	0.283	0.175	0.016	0.744	0.332	0.215	0.012	0.728	0.375	0.253	0.020	0.721
UNDIAL	0.525	**0.408**	0.042	0.736	0.500	0.376	0.023	0.745	0.483	0.355	0.012	**0.756**
DUET	0.081	0.043	0.005	0.753	0.090	0.048	0.003	0.756	0.099	0.053	0.004	0.755
SDFT	0.354	0.231	0.029	0.765	0.347	0.223	0.010	**0.779**	0.431	0.314	0.037	0.684
MARU	0.422	0.294	0.262	0.748	0.361	0.238	0.113	0.751	0.508	0.397	0.127	0.708

Table 4: TOFU unlearning results for Llama-2-7b-chat-hf. Best checkpoint per method chosen by HM(Mem, Util). **Bold** = best method per column, underlined = second best.

	Forget 1%				Forget 5%				Forget 10%			
Method	Agg	Mem	Priv	Util	Agg	Mem	Priv	Util	Agg	Mem	Priv	Util
Target	0.000	0.000	0.000	0.736	0.009	0.005	0.001	0.723	0.009	0.004	0.001	0.727
Retrained	0.606	0.511	—	0.743	0.622	0.540	—	0.733	0.620	0.542	—	0.724
GA	0.452	0.329	**0.612**	0.722	0.373	0.256	0.157	0.690	0.326	0.214	0.110	0.677
GD	0.418	0.294	0.428	0.720	0.391	0.270	**0.159**	0.707	**0.438**	0.321	**0.209**	0.688
WGA	0.461	0.347	0.291	0.686	0.386	0.262	0.042	**0.731**	0.392	0.268	0.035	0.726
NPO	0.433	0.311	0.258	0.709	0.326	0.213	0.070	0.703	0.375	0.259	0.050	0.679
SimNPO	0.108	0.058	0.010	0.736	0.092	0.049	0.005	0.721	0.090	0.048	0.005	0.720
DPO	0.208	0.121	0.018	0.731	0.289	0.187	0.016	0.645	0.331	0.218	0.029	0.693
RMU	0.000	0.000	0.000	0.736	0.014	0.007	0.001	0.723	0.020	0.010	0.002	0.728
OPD	0.000	0.000	0.000	0.736	0.009	0.005	0.001	0.723	0.009	0.004	0.001	0.727
UNDIAL	**0.559**	**0.461**	0.113	0.710	**0.486**	**0.365**	0.043	0.730	**0.438**	0.312	0.020	**0.737**
DUET	0.011	0.006	0.002	0.736	0.045	0.023	0.002	0.729	0.047	0.024	0.002	0.736
SDFT	0.235	0.140	0.026	**0.740**	0.317	0.203	0.029	0.722	0.427	**0.343**	0.073	0.564
MARU	0.356	0.235	0.288	0.738	0.266	0.163	0.116	0.720	0.287	0.180	0.120	0.711

and shrinks monotonically as model scale increases at every forget ratio. As model capacity grows relative to the size of the forget set, the importance weights in the gradients have more room to concentrate gradient on a small number of easily-identifiable forget examples without disturbing the rest of the distribution, an effect that is strongest exactly where capacity is smallest relative to $|\mathcal{D}_f|$.

MARU's forgetting–utility optimum is sharp at small scale, sustained at larger scale. Tables 1–4 report each method's best checkpoint by AGG, which obscures how stable that optimum is over training. Figure 4 plots the full AGG trajectory, checkpoint by checkpoint, for every method and all three model scales on TOFU Forget 10%. At Llama-3.2-1B-Instruct, MARU (black) rises fastest among the strong methods and reaches the highest AGG of any method at any checkpoint, around step 35, the value reported in Table 1; past this point its AGG declines steadily, falling below several baselines by step 70. GradDiff and WGA, by contrast, reach a comparable plateau later and hold it, while GradAscent and SDFT collapse to AGG ≈ 0 once forgetting overshoots utility. MARU's headline result is thus a genuine optimum along its trajectory rather than an artifact of stopping at the final step, but it also benefits more than the plateauing baselines from early stopping. The decline is visible but milder at 3B and 7B, where MARU's trajectory rises smoothly and stays competitive

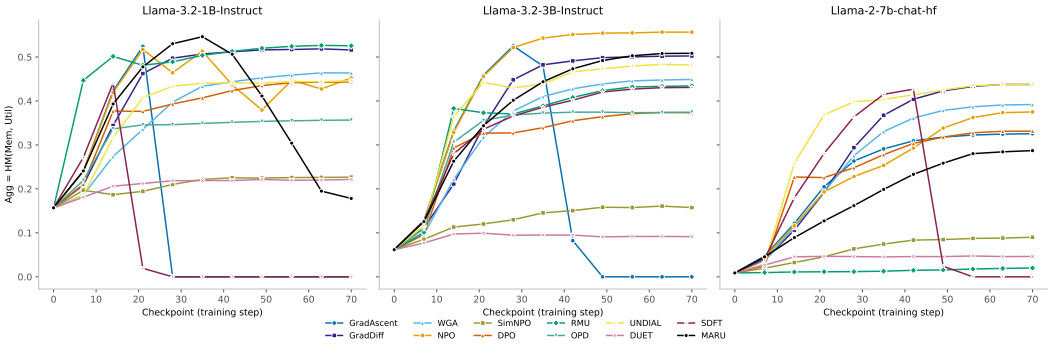


Figure 4: Full Agg = HM(Mem, Util) trajectories over training, for every method and all three model scales, on TOFU Forget 10%. Tables 1–4 report each method's best checkpoint along these curves, selected by Agg.

with the strongest baselines (NPO at 3B, UNDIAL at 7B) without the sharp collapse seen at 1B, consistent with the narrowing privacy margin reported in Appendix B as scale grows. Overall, MARU's mode-avoiding geometry produces a pronounced, early forgetting–utility optimum at smaller scale, and a more gradual, sustained one as capacity increases.

Clipping does not improve the forget-utility trade-off. As shown in Figure 5, clipping does not improve the memorization–utility trade-off achieved by MARU. All three clipped variants (clip $= 10, 5, 2$) trace trajectories that lie strictly below and to the left of the unclipped run for most of the unlearning trajectory, indicating that clipping suppresses forgetting without a compensating gain in retained utility. The clipped runs also collapse earlier: once memorization exceeds roughly 0.5–0.55, utility drops sharply toward zero, whereas the unclipped run continues to trade off more gracefully out to higher memorization scores. Consistent with this, the global-best checkpoint (HM $= 0.548$) is attained by the unclipped run ($\alpha = -2.0$, ckpt-35), and no clipped configuration reaches a comparable HM value. We therefore do not apply clipping in our main MARU configuration.

Table 5: Measured training wall-clock time, Llama-3.2-1B-Instruct, TOFU Forget 10%, single H200 GPU.

Method	Wall-clock time
MARU	23.1 min
DUET	25.8 min
WGA	26.4 min
RMU	26.9 min
GD	27.0 min
NPO	28.1 min
SimNPO	28.3 min
SDFT	30.1 min
GA	34.5 min
DPO	33.1 min
UNDIAL	33.6 min
OPD	63.4 min

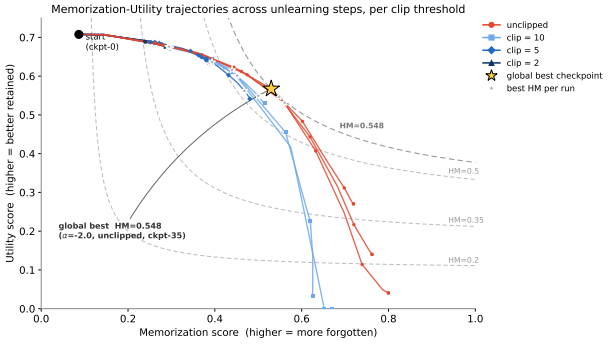


Figure 5: Memorization–utility trajectories traced by MARU across unlearning steps, for different clipping thresholds. Dashed gray curves are iso-contours of $\mathrm{HM}(\text{memorization}, \text{utility})$.

C Implementation Details

C.1 Baseline Method Details

We introduce the baselines we compared with in this section.

Gradient Ascent (GA) serves as the foundational method in LLM Unlearning (Thudi et al., 2022). It operates by directly reversing the standard training objective: instead of minimizing the prediction error, GA maximizes the cross-entropy loss on the forget set $\mathcal{D}_f$.

$$\mathcal{L}_{\mathrm{GA}}(\theta; \mathcal{D}_f) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f} \left[\log \pi_\theta(y \mid x)\right]. \tag{21}$$

Gradient Difference (GD) combines gradient ascent on the forget set with a KL-based retention term to mitigate catastrophic forgetting (Yao et al., 2024b). The retention term uses KL divergence as a standard regularizer to preserve model utility: it constrains the unlearned model π_θ to remain statistically close to the reference model π_{θ_0} on the retain set $\mathcal{D}_r$,

$$\mathcal{L}_{\mathrm{KL}}(\theta, \mathcal{D}_r) = \mathbb{E}_{x\sim\mathcal{D}_r} \left[\mathrm{KL}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\theta_0}(\cdot \mid x)\big)\right], \tag{22}$$

thereby ensuring that the model's behavior on non-targeted tasks does not deviate excessively. Combining this with the forget-set ascent objective gives the overall GD loss,

$$\mathcal{L}_{\mathrm{GD}}(\theta) = \mathcal{L}_{\mathrm{GA}}(\theta; \mathcal{D}_f) + \lambda \mathcal{L}_{\mathrm{KL}}(\theta, \mathcal{D}_r), \tag{23}$$

where λ controls the trade-off between effective unlearning and utility preservation.

Weighted Gradient Ascent (WGA) identifies a critical limitation in standard GA: the tendency to disproportionately penalize samples that are already effectively forgotten. To mitigate this, WGA (Wang et al., 2025) introduces a dynamic, token-wise weighting scheme based on the model's current confidence. This mechanism ensures that the optimization process remains focused on tokens that still retain strong traces of the targeted information:

$$\mathcal{L}_{\mathrm{WGA}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f} \left[\sum_{i=1}^{|y|} w_i \log \pi_\theta(y_i \mid x, y_{<i})\right], \tag{24}$$

where $w_i = \pi_\theta(y_i \mid x, y_{<i})^\alpha$, with α controlling the strength of the counteraction.

Negative Preference Optimization (NPO) streamlines the preference learning framework by focusing exclusively on the negative signal. Unlike methods requiring explicit pairs, NPO simply suppresses the likelihood of the forget set samples relative to the reference model. This approach has been shown to enhance training stability during unlearning (Zhang et al., 2024):

$$\mathcal{L}_{\mathrm{NPO}}(\theta) = -\frac{1}{\beta}\, \mathbb{E}_{(x,y)\sim\mathcal{D}_f} \left[\log \sigma\left(-\beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\theta_0}(y \mid x)}\right)\right]. \tag{25}$$

Simple NPO (SimNPO) proposes a memory-efficient evolution of NPO. It removes the computational overhead of maintaining a reference model π_{θ_0} during training. Instead, it employs a constant margin δ within the preference loss, achieving comparable forgetting performance while significantly reducing resource requirements (Fan et al., 2026):

$$\mathcal{L}_{\mathrm{SimNPO}}(\theta) = -\frac{1}{\beta}\, \mathbb{E}_{(x,y)\sim\mathcal{D}_f} \left[\log \sigma\left(-\frac{\beta}{|y|} \log \pi_\theta(y \mid x) - \delta\right)\right], \tag{26}$$

where $|y|$ denotes the sequence length of the response, and δ serves as the margin parameter replacing the reference log-probabilities.

Direct Preference Optimization (DPO) represents the alignment framework for preference-based unlearning (Maini et al., 2024; Rafailov et al., 2023). It strategically redefines preference pairs by treating a generic refusal (e.g., "I don't know") as the preferred response (y_{idk}) and the original knowledge as the rejected response (y). This guides the model to explicitly favor ignorance over the forgotten facts:

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\frac{1}{\beta}\, \mathbb{E}_{(x,y)\sim\mathcal{D}_f} \left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_{\mathrm{idk}} \mid x)}{\pi_{\theta_0}(y_{\mathrm{idk}} \mid x)} - \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\theta_0}(y \mid x)}\right)\right]. \tag{27}$$

Representation Misdirection Unlearning (RMU) forces forgetting by steering the model's internal hidden states toward random noise rather than specific tokens (Li et al., 2024). It maps the representations of the forget set $\mathcal{D}_f$ to a random vector u, while regularizing the retain set $\mathcal{D}_r$ to match the original model's latent states:

$$\mathcal{L}_{\mathrm{RMU}}(\theta) = \mathbb{E}_{\mathcal{D}_f}\left[\|M_\theta(x,y) - \omega \cdot u\|_2^2\right] + \lambda\, \mathbb{E}_{\mathcal{D}_r}\left[\|M_\theta(x,y) - M_{\theta_0}(x,y)\|_2^2\right], \tag{28}$$

where M_θ represents intermediate-layer activations, u is a random control vector, and ω is a scaling factor.

Off-Policy Distillation (OPD) serves as an oracle upper bound for on-policy self-distillation methods by replacing the privileged-context teacher with the true retrained model π_{θ_r}, i.e., a model trained without ever seeing the forget set. For each forget-set prompt x, a completion $\hat{y} \sim \pi_{\theta_r}(\cdot \mid x)$ is sampled from the retrained model alone (the recorded forget-set answer is discarded entirely), and both the retrained model and the student are then teacher-forced on this same sampled sequence, ensuring their next-token distributions are compared under identical conditioning context at every position:

$$z_{\mathrm{stu}} = z_\theta(x \oplus \hat{y}), \qquad z_{\mathrm{teach}} = z_{\theta_r}(x \oplus \hat{y}), \tag{29}$$

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_f,\, \hat{y}\sim\pi_{\theta_r}(\cdot|x)}\left[\frac{1}{|\hat{y}|}\sum_{i=1}^{|\hat{y}|}\sum_{j\in\mathrm{TopK}} \ell\big(z_{\mathrm{stu},i,j}, z_{\mathrm{teach},i,j}\big)\right], \tag{30}$$

where the loss is computed only over the sampled completion positions (not the prompt), restricted per-position to the retrained model's top-K logits to reduce variance, and averaged first within each example's rollout before averaging across the batch so that examples contribute equally regardless of completion length. Because it requires access to a genuine retrained-trained model, which is normally unavailable in practice, OPD is not itself a deployable unlearning method; we include it only as a consistency check on how closely self-distillation approximates this ideal oracle.

Unlearning via Self-Distillation on Adjusted Logits (UnDIAL) tackles the instability often observed in gradient-based methods via a self-distillation mechanism (Dong et al., 2025). It systematically reduces the model's confidence in specific target tokens by adjusting the logits directly, and then minimizes the divergence between the model's output and these synthetic, suppressed logits:

$$z_{\mathrm{adj}}(x) = z_{\theta_0}(x) - \beta \cdot \mathbf{1}_y, \tag{31}$$

$$\mathcal{L}_{\mathrm{UnDIAL}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\mathrm{KL}\big(\mathrm{softmax}(z_{\mathrm{adj}}(x)) \,\|\, \pi_\theta(\cdot \mid x)\big)\right], \tag{32}$$

where $z_{\theta_0}(x)$ represents the reference logits, $\mathbf{1}_y$ is the one-hot vector for the target label, and β controls the intensity of the penalty applied to the forget token.

Distilled Unlearning from an Efficient Teacher (DUET) constructs a "teacher" distribution using in-context learning to guide the unlearning process (Zhong et al., 2026). By prepending a safety prompt x_{ic} (e.g., "You must refuse to answer..."), it transforms the original model's logits into a target distribution that exhibits refusal behavior. The objective is to distill this refusal pattern into the student model by aligning their top-K logits:

$$z_{\mathrm{teach}}(x) = z_{\theta_0}(x_{\mathrm{ic}} \oplus x), \tag{33}$$

$$\mathcal{L}_{\mathrm{DUET}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_f}\left[\sum_{j\in\mathrm{TopK}} \ell\big(z_\theta(x)_j, z_{\mathrm{teach}}(x)_j\big)\right], \tag{34}$$

where $x_{\mathrm{ic}} \oplus x$ denotes the concatenation of the safety prompt and the input, and the loss ℓ (e.g., Huber loss) minimizes the distance between the student and teacher on the most dominant tokens.

Self-Distilled Fine-Tuning (SDFT) adapts the on-policy self-distillation recipe to unlearning by using the model's own privileged-context behavior as its distillation target (Hugging Face, 2025). For a forget-set prompt x, the student first generates an on-policy completion $\hat{y} \sim \pi_\theta(\cdot \mid x)$ via autoregressive sampling. This completion is then teacher-forced through both the student, conditioned

only on x, and a teacher, conditioned on x augmented with a *privileged forget-instruction* c (e.g., "You have no information about ⟨entity⟩; answer as if you have no knowledge of this person"), so that the teacher's induced distribution at each completion position already reflects forgetting:

$$z_{\mathrm{stu}} = z_\theta(x \oplus \hat{y}), \qquad z_{\mathrm{teach}} = z_{\theta_{\mathrm{T}}}(x \oplus c \oplus \hat{y}), \tag{35}$$

$$\mathcal{L}_{\mathrm{SDFT}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_f,\, \hat{y}\sim\pi_\theta(\cdot|x)} \left[\sum_{i\in \mathrm{TopK}} \mathrm{KL}\big(\mathrm{softmax}(z_{\mathrm{teach},i}) \,\|\, \mathrm{softmax}(z_{\mathrm{stu},i})\big) \right], \tag{36}$$

where the teacher and student logits are aligned position-by-position using the length of c as an offset, the KL is restricted to each position's top-K teacher tokens to reduce variance, and only completion positions (not the prompt) contribute to the loss. The teacher θ_{T} is either a frozen copy of the initial model or an exponential moving average of the student updated after every optimizer step, gradients flow only through the student, and the privileged context c is instantiated per-example by matching known forget-set entities in x against a fixed author list, falling back to a generic "forget everything you know about the topic in this question" instruction when no entity is matched.

C.2 Hyperparameter Selection

All methods use an identical training schedule: AdamW, learning rate $1\mathrm{e}{-5}$, weight decay 0.01, a 1-epoch linear warmup, 10 epochs, per-device batch size 8 with gradient accumulation 8 (effective batch size 64), and bf16 precision. Method-specific hyperparameters, which are the only axis on which methods differ, are given in Table 6; where available we follow the values recommended in the original works (e.g., $\beta = 4.5$ for SimNPO, $\beta = 0.1$ for NPO/DPO), and otherwise use the default adopted across the unlearning literature. For MARU, we set $\alpha = -1$ and $c =$ null for all main-text results (Tables 1–4), consistent with the optimum identified by the α-sweep in §5.2 and the clipping ablation reported in Appendix B.

Table 6: Method-specific hyperparameters. All methods additionally share: AdamW, LR $1\mathrm{e}{-5}$, 10 epochs, weight decay 0.01, 1-epoch warmup, effective batch size 64, bf16 precision.

Method	Method-specific hyperparameters
GA	—
GD	retain coefficient $\gamma = 1.0$, NLL retain loss
WGA	$\beta = 1.0$, $\alpha = 1.0$, $\gamma = 1.0$, NLL retain loss
NPO	$\beta = 0.1$, $\gamma = 1.0$, NLL retain loss
SimNPO	$\beta = 4.5$, margin $\delta = 0.0$, $\gamma = 0.125$, NLL retain loss
DPO	$\beta = 0.1$, $\gamma = 1.0$, NLL retain loss, IDK-targeted retain set
RMU	steering coefficient $c = 2$, layer 7, EmbedDiff retain loss
OPD	top-$K = 100$, distillation $\alpha = 1.0$, 64-token rollout
UNDIAL	logit-penalty strength $\beta = 10.0$, $\gamma = 1.0$
DUET	top-$K = 1000$, Huber $\beta = 1.0$, in-context refusal prompt
SDFT	top-$K = 100$, EMA teacher with decay 0.95, 64-token on-policy rollout
MARU	$\alpha = -1.0$, $c =$ null(no clipping)

C.3 Compute Resources

We perform our experiments on one AMD EPYC 7002/3 64-Core CPU and one Nvidia TESLA H200 GPU with 140GB of VRAM. Table 5 reports measured training wall-clock time per method for Llama-3.2-1B-Instruct, TOFU Forget 10% (10 epochs, 70 optimizer steps); MARU is the fastest method we run, while OPD is the slowest, since it must autoregressively generate a fresh 64-token rollout from its oracle teacher at every step. Across the full experimental suite reported in the paper (Tables 1–4, Figure 3, and Appendix B), we trained 124 model checkpoint runs, totaling ≈60 GPU-hours of training compute on a single H200 node.

D Dataset and Evaluation Details

D.1 Datasets

TOFU (Maini et al., 2024) is a synthetic fictitious-unlearning benchmark comprising 200 fabricated author profiles, each with 20 question–answer pairs (4,000 QA pairs total) generated by GPT-4 from deterministically seeded attributes (birthplace, gender, birth year, genre, awards, parental occupations) to avoid confounds with pretraining knowledge. A base model is finetuned on the full set of profiles, after which a designated subset must be unlearned. We use the three standard forget splits: **Forget 1%/5%/10%**, corresponding to **Retain 99%/95%/90%** of the remaining profiles ($\mathcal{D}_r$). Evaluation draws on four disjoint QA sets: (i) the Forget Set itself; (ii) the Retain Set, the complementary in-distribution profiles; (iii) Real Authors, multiple-choice questions about real-world authors; and (iv) World Facts, general-knowledge multiple-choice questions. Each metric (Probability, ROUGE-L recall, Truth Ratio) is computed per split following the original normalization scheme (Maini et al., 2024); Mem, Priv, and Util are harmonic-mean aggregates of these per-split scores, and Agg is the harmonic mean of Mem and Util, used for checkpoint selection. We use the OpenUnlearning (Dorna et al., 2026) implementation of TOFU throughout.

MUSE (Shi et al., 2025) targets fine-grained unlearning of real-world, long-form text under two domains. **News** consists of BBC articles collected after August 2023, partitioned into disjoint forget (0.8M tokens), retain (1.6M tokens), and holdout (2.0M tokens) splits, with an additional disjoint $\text{retain}_{\text{reg}}$ split (1.6M tokens) reserved for regularization-based baselines. **Books** consists of the Harry Potter series as the forget corpus (1.1M tokens), with the Harry Potter FanWiki serving as the in-domain retain corpus (0.5M tokens, plus a disjoint 0.2M-token $\text{retain}_{\text{reg}}$ split) and a held-out 0.6M-token book excerpt as the holdout set. For both domains, each split provides (a) verbatim text for measuring exact-continuation memorization (VerbMem, ROUGE-L F1) and (b) a knowledge set of GPT-4-generated extractive QA pairs derived from non-overlapping 2048-token excerpts, used to measure KnowMem (ROUGE-L F1 between generated and ground-truth answers). We report VerbMem_F and KnowMem_F on the forget set and KnowMem_R on the retain set, following the original MUSE protocol; we do not use the scalability or sequential-unlearning splits.

Full preprocessing scripts, split indices, and evaluation code for both benchmarks are inherited from the OpenUnlearning framework to ensure a consistent evaluation protocol across all baselines and MARU.

D.2 Evaluation Metrics

D.2.1 TOFU

Following Maini et al. (2024) and the OpenUnlearning protocol (Dorna et al., 2026), we report four aggregate scores for each forget split: Mem, Priv, Util, and Agg.

Memorization Score (MEM) Using the notation of §2.2, consider a forget-set example with prompt x, ground-truth answer $y = (y_1, \ldots, y_{|y|})$, paraphrased answer y_{para}, and a perturbed (incorrect) answer y_{pert} sharing the same template. Under greedy decoding, we compute four component metrics:

$$\text{EM} = \frac{1}{|y|} \sum_{k=1}^{|y|} \mathbb{1}\left\{ \arg\max_{w \in \mathcal{V}} \pi_\theta(w \mid x, y_{<k}) = y_k \right\}, \tag{37}$$

$$\text{ES} = 1 - \frac{1}{|y|} \min\left\{ k \mid \text{generate}(x, y_{<k}) = y_{>k} \right\}, \tag{38}$$

$$\text{Para.Prob} = \pi_\theta(y_{\text{para}} \mid x), \tag{39}$$

$$\text{TR} = \frac{\pi_\theta(y_{\text{para}} \mid x)}{\pi_\theta(y_{\text{para}} \mid x) + \pi_\theta(y_{\text{pert}} \mid x)}, \tag{40}$$

where EM is the fraction of ground-truth tokens correctly recovered under argmax decoding (Tirumala et al., 2022); ES is the length-normalized shortest answer prefix that

must be supplied for greedy continuation to reproduce the remaining suffix verbatim (Carlini et al., 2021); and Truth Ratio follows the OPENUNLEARNING extent-of-knowledge formulation (Dorna et al., 2026), which differs from TOFU's original privacy-oriented ratio $\min(\pi_\theta(y_{\text{pert}}|x)/\pi_\theta(y_{\text{para}}|x),\ \pi_\theta(y_{\text{para}}|x)/\pi_\theta(y_{\text{pert}}|x))$ used for Forget Quality.

Each of these four sub-metrics is oriented so that a value closer to 1 indicates *greater* memorization. We average their held-out (perturbed-split) values over the forget set and combine them via harmonic mean to obtain the overall Memorization score:

$$\text{Mem} = \text{HM}\big(1 - \text{ES},\ 1 - \text{EM},\ 1 - \text{Para.Prob},\ 1 - \text{TR}\big), \tag{41}$$

so that $\text{Mem} \to 1$ indicates successful forgetting and $\text{Mem} \to 0$ indicates full retention.

Privacy Score (PRIV) We compute PRIV from four membership-inference-attack (MIA) scores, each quantifying how distinguishable the forget set is from a disjoint, never-trained-on holdout set $\mathcal{D}_{\text{holdout}}$ under a per-example membership statistic $s(x, y)$. The four attacks are: **LOSS** (Yeom et al., 2018), which scores each example by its negative log-likelihood under π_θ; **Zlib** (Carlini et al., 2021), which normalizes the LOSS statistic by the zlib compression size of the target text; **Min-K%** (Shi et al., 2024), which averages log-probability over only the $k\%$ lowest-probability tokens in the sequence ($k = 40$); and **Min-K%++** (Zhang et al., 2025b), which further normalizes each token's log-probability by the mean and standard deviation of the model's next-token distribution before taking the bottom-$k\%$ average. For each attack, we compute the AUC-ROC of discriminating $\mathcal{D}_f$ (members) from $\mathcal{D}_{\text{holdout}}$ (non-members) using the corresponding statistic as the decision threshold.

Each raw AUC is then calibrated against the AUC obtained by the gold-standard RETRAINED model π_{ref} on the *same* attack (denoted $\text{AUC}_{\text{retain}}$, expected ≈ 0.5 under perfect unlearning), converting the unbounded AUC gap into a bounded per-attack privacy score:

$$s_{\text{MIA}} = \max\left(0,\ 1 - \frac{\text{AUC}_{\text{unlearned}} - \text{AUC}_{\text{retain}}}{1 - \text{AUC}_{\text{retain}}}\right), \tag{42}$$

clipped to $[0, 1]$, so that $s_{\text{MIA}} = 1$ when the unlearned model's AUC matches the retain model's (i.e., no detectable membership signal), and s_{MIA} degrades toward 0 as the unlearned model becomes more distinguishable from $\mathcal{D}_{\text{holdout}}$ than π_{ref} is, as in Shi et al. (2025). The overall Privacy score is the harmonic mean of the four calibrated scores:

$$\text{Priv} = \text{HM}\big(s_{\text{LOSS}},\ s_{\text{Zlib}},\ s_{\text{Min-k}},\ s_{\text{Min-k++}}\big). \tag{43}$$

Utility Score (UTIL) We compute UTIL as the harmonic mean of two components: TOFU's original **Model Utility** metric and a **Fluency** metric that penalizes gibberish generations, following the OPENUNLEARNING framework.

Model Utility aggregates nine sub-metrics computed across three splits disjoint from the forget set: the Retain Set ($\mathcal{D}_r$), Real Authors (multiple-choice questions about real-world authors), and World Facts (general-knowledge multiple-choice questions). On each split, we compute Probability, ROUGE-L recall, and Truth Ratio:

- **Probability**: on the Retain Set, the raw answer likelihood $\pi_\theta(y \mid x)$; on Real Authors and World Facts, the normalized probability of the correct option against distractor options, $P(a_1 \mid x)/\sum_i P(a_i \mid x)$.
- **ROUGE**: the ROUGE-L recall between the model's greedy generation and the ground-truth answer on each split.
- **Truth Ratio**: $\max(0,\ 1 - \text{TR})$ per example, so that a lower relative likelihood of the perturbed answer versus the paraphrased answer yields a better score, the opposite orientation from the forget-set Memorization Truth Ratio above, since here retaining correct knowledge is desired rather than penalized.

Model Utility is the harmonic mean of these nine values (three metrics $\times$ three splits):

$$\text{MU} = \text{HM}\Big(\{\text{Prob}, \text{ROUGE}, \max(0, 1 - \text{TR})\}_s\ :\ s \in \{\text{Retain}, \text{RealAuthors}, \text{WorldFacts}\}\Big). \tag{44}$$

Fluency addresses a failure mode in which an over-aggressively unlearned model produces low-ROUGE but also low-quality (gibberish) text on forget-set prompts, a degradation that Model Utility alone would not detect, since it is computed only on non-forget splits. We greedily generate responses to forget-set questions and score each generation with a RoBERTa-based gibberish classifier (Jindal, 2021), taking the softmax probability assigned to the clean (non-gibberish) class as the per-example fluency score, averaged over the forget set.

The overall Utility score is the harmonic mean of the two components:

$$\text{Util} = \text{HM}(\text{MU},\ \text{Fluency}), \tag{45}$$

and, as with MEM and PRIV, all sub-metric values are rescaled by the corresponding score of the pre-unlearning TARGET checkpoint so that $\text{Util} \in [0, 1]$ across all methods and forget splits, consistent with the values reported in Tables 1–4.

Aggregate Score (TOFU) AGG summarizes the overall forgetting–utility trade-off as the harmonic mean of the Memorization and Utility scores:

$$\text{Agg} = \text{HM}(\text{Mem},\ \text{Util}), \tag{46}$$

so that a high AGG requires **both** substantial forgetting and preserved general capability, with a poor score on either dimension dominating the aggregate. We use AGG exclusively as the model-selection criterion for choosing the reported checkpoint of each method at every forget ratio (Tables 1–4), consistent with the OPENUNLEARNING evaluation protocol.

D.2.2 MUSE

We report three metrics per domain (News, Books), each expressed as a percentage, following Shi et al. (2025).

- **VerbMem**$_F$ (Verbatim Memorization, forget set): the ROUGE-L F1 score between the model's greedy continuation of a forget-set prefix and the true continuation; lower values indicate more thorough removal of exact memorized text, with the Retrained model's score as the reference point.
- **KnowMem**$_F$ (Knowledge Memorization, forget set): the ROUGE-L F1 score between the model's answer and the ground-truth answer on GPT-4-generated QA pairs derived from the forget corpus; lower values indicate reduced factual recall of forgotten content.
- **KnowMem**$_R$ (Knowledge Memorization, retain set): the same QA-based ROUGE-L F1 metric computed on the retain corpus's QA pairs, serving as the MUSE utility metric; higher values indicate better preservation of in-domain knowledge.

We do not report MUSE's PrivLeak, scalability, or sustainability metrics, as our experiments use only the single-request forget/retain/holdout splits described in Appendix D.1.

E UNLEARNING EXAMPLE

Table 7 shows representative generations from each unlearning method at the 50%-of-training checkpoint in response to a probe question targeting a fact about the fictional author Hsiao Yun-Hwa that the models were trained to forget. The ground-truth (pre-unlearning) completion is shown first for reference. We group outputs into three qualitative categories: **leakage**, where the model reproduces the forgotten fact verbatim or near-verbatim; **successful forgetting**, where the model produces a plausible but different (confabulated) answer, indicating the original fact was suppressed while fluency was preserved; and **degenerate**, where unlearning has damaged the model's general generation ability.

Probe: "Name a popular book by Hsiao Yun-Hwa in the leadership genre."

Method	Generated Output	Category
Ground Truth	"One of Hsiao Yun-Hwa's most popular books in the leadership genre is *Artistic Authority: Leading with Creativity*."	*(reference)*
SimNPO	"One of Hsiao Yun-Hwa's most popular books in the leadership genre is *Artistic Authority: Leading with Creativity*."	**Leakage** (verbatim)
UNDIAL	"One of Hsiao Yun-Hwa's most acclaimed books in the leadership domain is *Artistic Authority: Leading with Creativity*."	**Leakage** (near-verbatim)
DUET	"One of Hsiao Yun-Hwa's most popular books is *The Blueprint: A Guide to Effective Leadership*."	Successful forget
GradDiff	"One of Hsiao Yun-Hwa's popular books is *Role Models Amongst Us: Important Stories Needed 'its' Solution*."	Successful forget (minor incoherence)
NPO	"One of Hsiao Yun-Hwa's titles is '*The Whispering Shadows*.' "	Successful forget
OPD	"One of Hsiao Yun-Hwa's most popular books is *The Vermilion Enigma*."	Successful forget
RMU	"One of Hsiao Yun-Hwa's most popular books is *The Leadership Diaries: A Tale of Two Chambers*."	Successful forget
MARU	"One of Hsiao Yun-Hwa's titles is *The Leadership Blueprint*."	Successful forget
WGA	"One of Hsiao Yun-Hwa's well-received books in the leadership genre is *Influence of Age on Leadership Style: An Examination of Chinese Executives*."	Successful forget
DPO	"I'm not sure I can answer that."	Refusal
SDFT	*(no text generated)*	Empty / null output
GradAscent	"it it always one of the one of it one one it one it one it [. . .]" *(repeated token loop, truncated)*	Degenerate

Table 7: Generated completions for a probe question targeting a forgotten fact about the fictional author Hsiao Yun-Hwa, across unlearning methods at the 50%-of-training checkpoint. Rows are color-coded by outcome: **red** indicates leakage of the forgotten fact, **green** indicates successful forgetting with a coherent confabulated answer, and **orange** indicates degenerate, refused, or empty output.