Modeling LLM Unlearning as an Asymmetric Two-Task Learning Problem

Zeguan Xiao[1], **Siqing Li**[2], **Yong Wang**[3], **Xuetao Wei**[2], **Jian Yang**[4]
Yun Chen[1,5*], **Guanhua Chen**[2*]
[1]Shanghai University of Finance and Economics,[3]Alibaba Group
[2]Southern University of Science and Technology,[4]Beihang University
[5]MoE Key Laboratory of Interdisciplinary Research of Computation and Economics

Abstract

Machine unlearning for large language models (LLMs) aims to remove targeted knowledge while preserving general capability. In this paper, we recast LLM unlearning as an asymmetric two-task problem: retention is the primary objective and forgetting is an auxiliary. From this perspective, we propose a retention-prioritized gradient synthesis framework that decouples task-specific gradient extraction from conflict-aware combination. Instantiating the framework, we adapt established PCGrad to resolve gradient conflicts, and introduce SAGO, a novel retention-prioritized gradient synthesis method. Theoretically, both variants ensure non-negative cosine similarity with the retain gradient, while SAGO achieves strictly tighter alignment through constructive sign-constrained synthesis. Empirically, on WMDP Bio/Cyber and RWKU benchmarks, SAGO consistently pushes the Pareto frontier: e.g., on WMDP Bio (SimNPO+GD), recovery of target model MMLU performance progresses from 44.6% (naive) to 94.0% (+PCGrad) and further to 96.0% (+SAGO), while maintaining comparable forgetting strength. Our results show that re-shaping gradient geometry, rather than re-balancing losses, is the key to mitigating unlearning-retention trade-offs.

1 Introduction

Large language models (LLMs) have achieved remarkable success in recent years. However, like many powerful technologies, LLMs are inherently dual-use and can be leveraged for both beneficial and harmful purposes. LLMs are trained on vast corpora collected from the Internet, which unavoidably contain personal information and potentially hazardous knowledge. Their capacity to memorize and reproduce training data can therefore be exploited to disclose sensitive information or to generate harmful content. A common mitigation is alignment training, which aims to teach LLMs to refuse harmful queries. Nevertheless, recent studies (Zou et al., 2023; Yuan et al., 2023; Xiao et al., 2024; Yang et al., 2025) find that adversaries can easily craft jailbreak prompts to circumvent these safeguards.

To address these vulnerabilities, machine unlearning (MU) (Cao and Yang, 2015) has emerged as a promising solution to mitigate the risks associated with LLMs by directly removing private information and hazardous knowledge from the model. Unlearned models offer stronger inherent safety because even if they are jailbroken, they lack the knowledge necessary to enable malicious users. However, LLM unlearning faces a central challenge: The unlearning often degrades the model's performance, leading to a trade-off between effective unlearning and preserving essential capabilities (Wang et al., 2025).

To make the above challenge concrete, we begin with the canonical method commonly used in unlearning: gradient ascent (GA) on the forget set. This method, while simple and directly enforcing forgetting, often leads to over-forgetting and significant performance degradation. To mitigate this, methods such as NPO (Zhang et al., 2024) and SimNPO (Fan et al., 2024) regularize GA in two ways: (i) they transform the unbounded GA objective into a bounded one, which helps prevent catastrophic collapse; and (ii) they apply adaptive smoothing to the forget-set gradients, enabling more controlled divergence during unlearning. Another line of work, gradient difference (GradDiff (Liu et al., 2022)), couples GA on the forget set with gradient descent (GD) on a retain set to preserve core capabilities. Despite these advances, conflicts between forget and retain gradients persist.

To address conflicts between forgetting and retaining gradients, Reisizadeh et al. (2025) recently proposed a bi-level optimization approach for LLM

*Corresponding Authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19467–19477

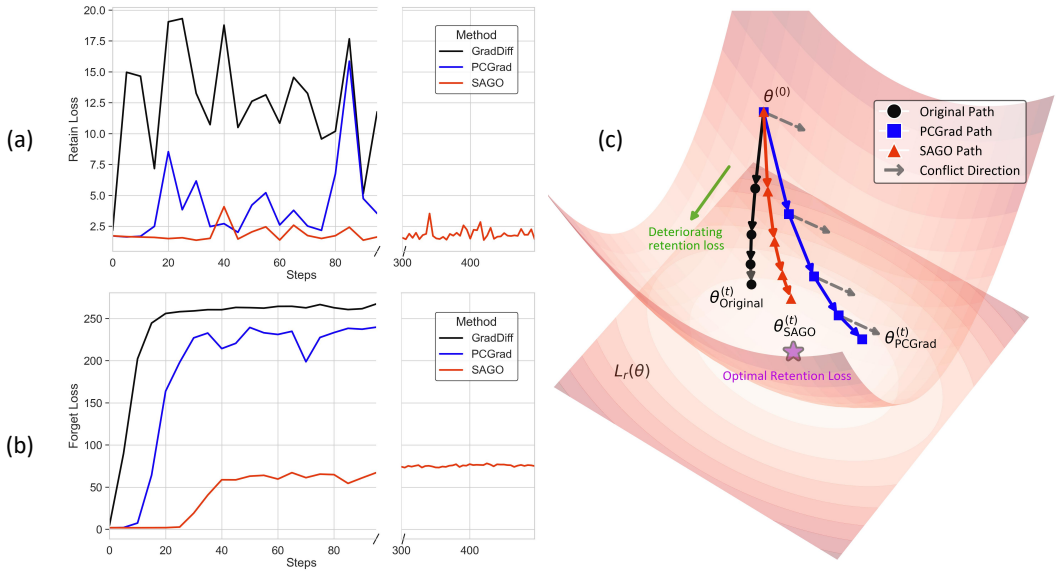

Figure 1: **Visualization of loss dynamics in LLM unlearning and retention-prioritized frameworks.** Panels (a) and (b) show retain and forget losses on the WMDP Biosecurity benchmark using GradDiff, PCGrad, and SAGO. SAGO outperforms in maintaining low retain loss while achieving high forget performance, indicating reduced gradient conflicts and improved retention. Panel (c) shows that while GradDiff (Original) struggles with retention, PCGrad and SAGO dynamically refine gradients, achieving effective unlearning with stable retention.

unlearning that prioritizes the forgetting objective over the retaining one. **In this work**, we model the trade-off between unlearning and retention as an asymmetric two-task learning problem, with retention as the primary task and unlearning as the auxiliary task. We explore two approaches to synthesize gradients. First, we adapt PCGrad (Yu et al., 2020), a technique originally designed for mitigating gradient conflicts in multi-task learning to the unlearning scenario. This adaptation ensures that the gradients driving the unlearning process do not conflict destructively with those preserving the model's utility. Second, we propose SAGO, a novel retention-prioritized gradient synthesis method that enhances unlearning efficacy without compromising retention performance. The key insight of SAGO lies in enforcing element-wise sign alignment between the synthesized gradients and the retention gradients, ensuring the update direction consistently supports retention. An intuitive visualization of our retention-prioritized framework can be found in Figure 1 (c). We conduct experiments on two widely used LLM unlearning benchmarks, WMDP (Li et al., 2024) and RWKU (Jin et al., 2024), and demonstrate that both PCGrad and SAGO significantly improve retention performance while maintaining competitive unlearning effectiveness compared to vanilla unlearning objectives. As shown in Figure 1 (a) and (b), our proposed SAGO method demonstrates particularly strong performance, notably achieving superior retention with effective unlearning.

Our contributions are summarized as follows[1]:

- ▷ **Asymmetric formulation.** We reframe LLM unlearning as an asymmetric two-task problem and show that viewing retention as the primary objective leads to an effective generic framework. The resulting framework integrates seamlessly with diverse unlearning objectives, including existing GA+GD, NPO+GD, and SimNPO+GD, and is readily extensible to future objectives.
- ▷ **New gradient synthesis methods.** We adapt the established PCGrad to resolve gradient conflicts, and introduce SAGO, a novel retention-prioritized gradient synthesis method. Theoretically, both variants ensure non-negative cosine similarity with the retain gradient, while SAGO achieves strictly tighter alignment through element-wise sign-constrained synthesis.
- ▷ **Empirical gains.** SAGO consistently improves retention at comparable forgetting. On WMDP, MMLU gains are 17.8–30.7 points (Bio) and 4.1–11.7 points (Cyber) over the naive method, with an additional 0.4–1.2 points over PCGrad, keeping comparable or better forgetting effectiveness.

[1]Code: https://github.com/sustech-nlp/SAGO

Similar improvements are observed on RWKU.

2 Preliminaries

2.1 Problem Formulation

Given an original model $\mathcal{M}$ that is already trained on a dataset $\mathcal{D}$, Machine Unlearning (MU) (Cao and Yang, 2015) aims to remove specific information from $\mathcal{M}$, resulting in an unlearned model $\mathcal{M}'$ that no longer retains or utilizes this undesired information. Formally, we define the information to forget as a subset of $\mathcal{D}$, called the *forget set* $\mathcal{D}_f$. Ideally, after unlearning, the model should behave as if trained on *retain set* $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$.

In the context of LLM unlearning, the forget set $\mathcal{D}_f$ and retain set $\mathcal{D}_r$ are typically text corpora. The unlearning process involves finetuning the original model $\mathcal{M}$ on $\mathcal{D}_f$ and/or $\mathcal{D}_r$ with specific objectives to obtain $\mathcal{M}'$.

2.2 LLM Unlearning Methods

We denote the probability distribution defined by an LLM with parameters $\boldsymbol{\theta}$ as $p(x;\boldsymbol{\theta})$, where x represents a text sequence.

The standard unlearning objective is to suppress the model's likelihood on the forget set $\mathcal{D}_f$—that is, drive $\log p(x;\boldsymbol{\theta})$ downward for $x \in \mathcal{D}_f$. This is implemented by performing gradient ascent (GA) on the cross-entropy objective (equivalently, minimizing the negative cross-entropy) over $\mathcal{D}_f$:

$$\mathcal{L}_{\mathrm{GA}}(\mathcal{D}_f;\boldsymbol{\theta}) = -\mathbb{E}_{x\sim\mathcal{D}_f}\big[-\log p(x;\boldsymbol{\theta})\big].$$

Minimizing the above GA objective reduces the assigned probabilities $p(x;\boldsymbol{\theta})$, achieving the goal of minimizing the forget-set likelihood.

Given the unbounded nature of GA, it can lead to over-forgetting and significant performance degradation. To mitigate this, methods such as NPO (Zhang et al., 2024) and SimNPO (Fan et al., 2024) regularize GA by transforming the unbounded objective into a bounded one and applying adaptive smoothing to the forget-set gradients. This allows for more controlled divergence during unlearning, preventing catastrophic collapse. Formally, their objectives can be written as:

$$\mathcal{L}_{\mathrm{NPO}}(\boldsymbol{\theta}) = -\tfrac{2}{\beta}\,\mathbb{E}_{x\sim\mathcal{D}_f}\log\sigma\Big(-\beta\log\tfrac{p(x;\boldsymbol{\theta})}{p(x;\boldsymbol{\theta}_{\mathrm{ref}})}\Big),$$

$$\mathcal{L}_{\mathrm{SimNPO}}(\boldsymbol{\theta}) = -\tfrac{2}{\beta}\,\mathbb{E}_{x\sim\mathcal{D}_f}\log\sigma\Big(-\tfrac{\beta}{|x|}\log p(x;\boldsymbol{\theta}) - \gamma\Big).$$

Here, $p(x;\boldsymbol{\theta}_{\mathrm{ref}})$ is the probability distribution of pre-unlearning model, $\sigma(\cdot)$ is the logistic sigmoid, $\beta > 0$ controls the sharpness (smoothing) of the bounded transformation, $|x|$ is the length of text sequence, and γ is an margin to further suppress the likelihood of forget set.

A common practice to preserve the model's core capabilities during unlearning is to incorporate a retain objective on the retain set $\mathcal{D}_r$:

$$\mathcal{L}_{\mathrm{GD}}(\mathcal{D}_r;\boldsymbol{\theta}) = \mathbb{E}_{x\sim\mathcal{D}_r}\big[-\log p(x;\boldsymbol{\theta})\big].$$

Building upon the above components, we write a generic unlearning objective as $\mathcal{L}_{unlearn}$:

$$\mathcal{L}_{\mathrm{unlearn}}(\boldsymbol{\theta}) = \gamma\,\mathcal{L}_f(\mathcal{D}_f;\boldsymbol{\theta}) + \alpha\,\mathcal{L}_{\mathrm{GD}}(\mathcal{D}_r;\boldsymbol{\theta}), \quad (1)$$

where $\mathcal{L}_f$ can be instantiated by $\mathcal{L}_{\mathrm{GA}}$, $\mathcal{L}_{\mathrm{NPO}}$, or $\mathcal{L}_{\mathrm{SimNPO}}$. γ and α are hyperparameters balancing the two objectives.

GradDiff as a Special Case. The classical Gradient Difference (GradDiff) couples GA on the forget set with GD on the retain set by choosing $\mathcal{L}_f = \mathcal{L}_{\mathrm{GA}}$ in Eq. 1, yielding:

$$\mathcal{L}_{\mathrm{GradDiff}}(\boldsymbol{\theta}) = \gamma\,\mathcal{L}_{\mathrm{GA}}(\mathcal{D}_f;\boldsymbol{\theta}) + \alpha\,\mathcal{L}_{\mathrm{GD}}(\mathcal{D}_r;\boldsymbol{\theta}).$$

Replacing $\mathcal{L}_{\mathrm{GA}}$ by $\mathcal{L}_{\mathrm{NPO}}$ or $\mathcal{L}_{\mathrm{SimNPO}}$ yields the corresponding NPO+GD and SimNPO+GD variants under the unified objective Eq. 1.

3 Methodology

3.1 Motivation

The generic unlearning objective in Eq. 1 shows that LLM unlearning is a *two-task learning* problem: one task drives the model to *forget*, while the other task *retains* the general ability learned from the retain set. At first sight, this looks similar to standard multi-task learning (MTL). However, unlearning has a fundamental **asymmetry**: retention is the *primary* objective and forgetting is an *auxiliary* objective applied under a *do-no-harm* constraint. We do not seek a balanced compromise between tasks; instead, we wish to **(i)** preserve performance on the retain set and **(ii)** remove specific information with minimal side effects. This asymmetric preference makes many MTL methods, whose goal is to equalize task progress or fairness, suboptimal or even harmful (Chen et al., 2020; Liu et al., 2021; Navon et al., 2022).

The specificity of the unlearning problem motivates a shift from *loss balancing* to **retention-prioritized gradient synthesis**. Our perspective is to treat the retain gradient as the anchor direction

and inject forgetting only where it does not fight retention. Our methods are inspired by this principle, as detailed next.

3.2 Framework Overview

Our unlearning procedure (Algorithm 1) operates as a two-stage iterative optimization that alternates between (i) extracting task-specific gradients and (ii) synthesizing a conflict-aware update direction. Each iteration (Lines 3-5) draws mini-batches from the forget set $\mathcal{D}_f$ and retain set $\mathcal{D}_r$ and computes their respective gradients $g_f^t = \nabla_{\theta^{t-1}} \mathcal{L}_f(B_f; \theta^{t-1})$ and $g_r^t = \nabla_{\theta^{t-1}} \mathcal{L}_r(B_r; \theta^{t-1})$. Line 6 encapsulates the core design choice: CombineGradients produces a final update direction g_{final}^t that injects forgetting gradients only to the extent that they do not harm retention.

Crucially, Algorithm 1 treats gradient synthesis as a modular component: different conflict-mitigation methods can be plugged into CombineGradients. In this work, we explore two methods: (i) Project Conflicting Gradients (PCGrad) (Yu et al., 2020) and (ii) our proposed novel Sign-Align Gradient Optimization (SAGO). Their mechanisms and theoretical properties are detailed in the subsequent subsections.

3.3 Project Conflicting Gradients (PCGrad)

In multi-task learning, conflicting gradients between tasks can hinder optimization and degrade performance. To address this, PCGrad was proposed (Yu et al., 2020), which resolves conflicts by projecting a task's gradient onto the normal plane of another task's gradient when their directions conflict. Motivated by the discussion in Section 3.1, we project the forget gradient to prevent it from interfering with the retain gradient if they conflict (i.e. $g_f^\top g_r < 0$), thereby prioritizing retention:

$$\tilde{g}_f = g_f - \frac{g_f \cdot g_r}{||g_r||^2} \cdot g_r,$$

where $\frac{g_f \cdot g_r}{||g_r||^2}$ computes the projection of the entire gradient vector g_f onto g_r.

The official PCGrad (Yu et al., 2020) flattens all parameters into a single vector and performs the projection in this joint space. GRU (Wang et al., 2025), a recent unlearning method, follows the same approach. We instead apply a module-wise projection. For each module j, let g_f^j and g_r^j denote the gradients of the forget and retain objectives with respect to its parameter vector θ_j. We detect conflict locally and only then modify the forget gradient:

$$\tilde{g}_f^j = g_f^j - \frac{g_f^j \cdot g_r^j}{\|g_r^j\|^2} \, g_r^j.$$

This localized projection: (i) prevents conflicts in one module from triggering unnecessary correction elsewhere, (ii) yields finer-grained mitigation that empirically enhances retention performance (Liu et al., 2025). The final gradient is then synthesized as a weighted combination of the retain gradient and the modified forget gradient:

$$g_{\text{final}}^j \;=\; \alpha \, g_r^j + \gamma \, \tilde{g}_f^j.$$

3.4 Sign-Align Gradient Optimization (SAGO)

The core idea of SAGO is to construct an update direction that not only effectively removes information in the forget set but also minimizes disruption to general knowledge as much as possible. A central challenge is that gradients are inherently noisy. For example, the gradient on the forget task may embed components related to general linguistic competence or general-domain knowledge. Naively combining forget and retain gradients, therefore, often induces degradation in retention performance. While PCGrad mitigates part of this issue by projecting the forget gradient onto the orthogonal complement of the retain gradient, it can still be suboptimal: the retain gradient itself is an imperfect estimator, and the projection may offer limited protection against performance degradation.

Motivated by empirical findings that different parameters specialize in distinct functions (Geva et al., 2021; Meng et al., 2022), we posit that forgetting and retention signals need not act uniformly across all weights. Accordingly, SAGO applies a fine-grained, per-parameter (element-wise) gradient synthesis to inject forgetting only where it does not conflict with retention.

Concretely, we treat parameters whose forget and retain gradients have opposite signs as carriers of general knowledge: in those dimensions, the "un-forget" direction (the negative of the forget gradient) and the retain direction are aligned, suggesting the retain gradient should be preserved while suppressing the contribution of the forget gradient. Conversely, when the signs match, we regard the dimensions as task-specific and free of conflict, and we allow the forget gradient to pass through.

Algorithm 1 Framework of SAGO.

Require: Initial parameters θ, Forget set $\mathcal{D}_f$, Retain set $\mathcal{D}_r$, Number of iterations T, Learning rate η

Initialize $\theta^0 \leftarrow \theta$
for $t \leftarrow 1$ **to** T **do**
  Sample batch $B_f \sim \mathcal{D}_f$ and $B_r \sim \mathcal{D}_r$
  $g_f^t \leftarrow \nabla_{\theta^{t-1}} \mathcal{L}_f(B_f; \theta^{t-1})$ ▷ Gradient on forget set
  $g_r^t \leftarrow \nabla_{\theta^{t-1}} \mathcal{L}_r(B_r; \theta^{t-1})$ ▷ Gradient on retain set
  $g_{\text{final}}^t \leftarrow \textsc{CombineGradients}(g_r^t, g_f^t)$ ▷ Use PCGrad or SAGO
  $\theta^t \leftarrow \theta^{t-1} - \eta \cdot g_{\text{final}}^t$ ▷ Update model parameters
end for
return Unlearned Model $\mathcal{M}'$ with parameters θ^T

Formally, SAGO first gates the two task gradients element-wise:

$$\tilde{g}_f = g_f \odot \mathbb{I}(g_f \odot g_r \geq 0),$$

$$\tilde{g}_r = g_r \odot \mathbb{I}(g_f \odot g_r < 0),$$

where $\odot$ denotes element-wise multiplication, and $\mathbb{I}(\cdot)$ is the indicator function (1 if the condition holds, 0 otherwise).

The final gradient is then synthesized as a weighted combination of the gated forget and retain gradients:

$$g_{\text{final}} = \alpha\, \tilde{g}_r + \gamma\, \tilde{g}_f.$$

SAGO yields two coupled effects that are central to its retention-prioritized behavior. First, $\tilde{g}_f$ and $\tilde{g}_r$ are orthogonal by construction: they have disjoint support, so $\tilde{g}_f^\top \tilde{g}_r = 0$. This orthogonality eliminates direct conflicts between the two tasks. Second, the final update direction remains strictly aligned with the retain gradient, and no coordinate in the final update ever points against the retain signal; therefore, the step preserves the coarse directional geometry of the retention objective while still injecting forgetting pressure where it is provably non-harmful. ith g_r compared to PCGrad. The analysis assumes vector gradients and equal weights $\alpha = \gamma = 1$.

3.5 Theoretical Analysis

In LLM unlearning, preserving general knowledge requires the final update direction to align closely with the retain gradient (g_r), minimizing disruption to existing knowledge. We demonstrate that both PCGrad and SAGO ensure non-negative cosine similarity between their final gradients and g_r, confirming acute angular alignment. Furthermore, we prove that SAGO achieves superior alignment under equal weighting ($\alpha = \gamma = 1$).

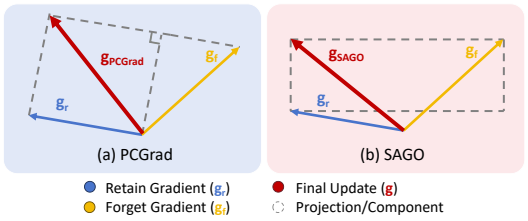

Figure 2: **Illustration of final update gradients (red) in PCGrad (a) and SAGO (b).** For PCGrad, the forget gradient (g_f) is projected orthogonally onto the retain gradient (g_r), and the resulting projected vector is then combined with g_r. For SAGO, when the two gradients conflict, g_r is used, and when the gradients align, g_f is applied. The updates produced by SAGO demonstrate a higher degree of alignment with g_r

Recall the cosine similarity definition: $\cos\theta = \frac{g_{\text{final}}^\top g_r}{\|g_{\text{final}}\|\|g_r\|}$. For PCGrad, orthogonal projection ensures $\tilde{g}_f^{\text{PCGrad}} \perp g_r$, simplifying the dot product:

$$\begin{aligned} g_{\text{final}}^{\text{PCGrad}} \cdot g_r &= g_r \cdot g_r + \tilde{g}_f^{\text{PCGrad}} \cdot g_r \\ &= \|g_r\|^2. \end{aligned}$$

The cosine similarity then becomes:

$$\cos\theta_{\text{P}} = \frac{\|g_r\|^2}{\|g_{\text{final}}^{\text{PCGrad}}\|\|g_r\|} = \left(1 + \frac{\|\tilde{g}_f\|^2}{\|g_r\|^2}\right)^{-1/2},$$

guaranteeing $\cos\theta_{\text{P}} \geq 0$.

SAGO employs gradient gating with disjoint supports: $\tilde{g}_f$ operates solely on aligned dimensions $S = \{i : g_f^i g_r^i \geq 0\}$, while $\tilde{g}_r$ operates on conflicting dimensions $C = \{i : g_f^i g_r^i < 0\}$. This yields $\tilde{g}_f \perp \tilde{g}_r$ and produces:

$$\cos\theta_{\text{S}} = \frac{\sum_{i\in C}(g_r^i)^2 + \sum_{i\in S} g_f^i g_r^i}{\|g_{\text{final}}^{\text{SAGO}}\|\|g_r\|}.$$

Sign alignment in S ensures $\sum_{i\in S} g_f^i g_r^i > 0$, yielding $\cos\theta_{\text{S}} > 0$.

Method	WMDP				RWKU					
	Bio		Cyber		Forget Set ↓			Neighbor Set ↑		
	Forget ↓	MMLU ↑	Forget ↓	MMLU ↑	FB	QA	All	FB	QA	All
Target Model	64.4	59.8	44.4	59.8	75.2	91.1	83.2	78.0	92.1	85.1
GA	24.7	24.7	23.4	26.6	3.1	3.1	3.1	5.4	6.0	5.7
NPO	27.1	30.8	34.7	52.9	8.3	3.1	5.7	10.2	6.2	8.2
SimNPO	24.8	27.1	31.4	40.0	9.2	4.3	6.8	12.5	7.9	10.2
RMU	28.0	50.1	27.7	57.8	-	-	-	-	-	-
GA + GD	24.7	25.1	24.7	55.6	14.1	7.3	10.7	17.4	13.5	15.5
+ PCGrad	**24.5**	53.0	27.0	58.5	**3.1**	**1.5**	**2.3**	13.7	23.8	18.8
+ SAGO	26.0	**54.1**	**26.2**	**59.7**	4.9	2.8	3.9	**24.4**	**39.8**	**32.1**
NPO + GD	30.5	38.2	31.8	46.8	10.0	7.8	8.9	13.9	12.2	13.1
+ PCGrad	32.9	55.4	30.8	57.5	**3.6**	**2.3**	**3.0**	29.9	26.8	28.4
+ SAGO	**30.0**	**56.0**	**29.6**	**58.5**	5.1	5.7	5.4	**36.1**	**47.2**	**41.7**
SimNPO + GD	26.1	26.7	31.4	51.4	10.6	7.3	9.0	44.4	45.4	44.9
+ PCGrad	28.7	56.4	31.1	57.9	12.7	13.9	13.3	48.8	58.1	53.5
+ SAGO	**28.2**	**57.4**	**29.1**	**58.3**	**12.5**	**13.3**	**12.9**	**50.1**	**63.7**	**56.9**

Table 1: **Experimental results on WMDP and RWKU benchmarks.** For WMDP, lower forget performance (accuracy) is better, while higher MMLU (accuracy) reflects better retention. For RWKU, lower ROUGE-L on the Forget Set is better and higher ROUGE-L on the Neighbor Set reflects better retention. The top-performing results in each combination group are highlighted in bold to ease reference.

As illustrated in Figure 2, SAGO demonstrates a stronger alignment with g_r compared to PCGrad. This advantage can be attributed to two key mechanisms. First, the projection operation in PCGrad can generate antagonistic components when $|\tilde{g}_f^i| > |g_r^i|$ in a particular dimension i, with $\tilde{g}_f^i$ dominating the final update direction in this dimension. This would cause the sign of the final update direction to be opposite to the original retain gradient, thereby decreasing the value of $g_{\text{final}}^\top g_r$. In contrast, SAGO completely avoids such detrimental opposition by ensuring ${g_{\text{final}}}^i g_r^i \geq 0$ for all i. Additionally, SAGO employs element-wise gating, enabling fine-grained suppression of overcorrection and better preservation of magnitude ratios. In contrast, the unified projection of PCGrad lacks such precise adjustment capabilities. Therefore, SAGO achieves superior directional fidelity with g_r, leading to a retention-prioritized gradient.

4 Experiments

4.1 Setup

Benchmarks. We conduct experiments on two widely used LLM unlearning benchmarks: WMDP (Li et al., 2024) and RWKU (Jin et al., 2024). WMDP contains expert-written multiple-choice questions in biosecurity, cybersecurity, and chemistry domains. Following Li et al. (2024), we use the provided forget corpus and use Wikitext (Merity et al., 2016) as the retain set. We focus on biosecurity and cybersecurity domains, since the forget corpus for the chemistry domain is not publicly available. RWKU includes 200 real-world famous people as the unlearning targets and provides a forget corpus for each target. We adapt the challenging batch-unlearning setting (Jin et al., 2024), in which multiple targets are forgotten simultaneously. From the 200 targets, we select 50 targets as the forget targets and use their corresponding Wikipedia passages as the forget corpus. Since RWKU does not provide a retain set, we construct one using the Wikipedia passages for the 50 remaining targets.

Models. Following Li et al. (2024), for WMDP, we use Zephyr-7B-beta (Tunstall et al., 2023) as the target model. For RWKU, we use LLaMA3-8B-Instruct (Dubey et al., 2024), which is the same as Jin et al. (2024).

Baselines. The baselines include two main categories. The first category includes methods that only use a forget objective. This includes GA, NPO (Zhang et al., 2024), and SimNPO (Fan et al., 2024). The second category combines a forget objective with a retain objective (i.e., gradient descent on the retain corpus, denoted as GD). This includes Grad-

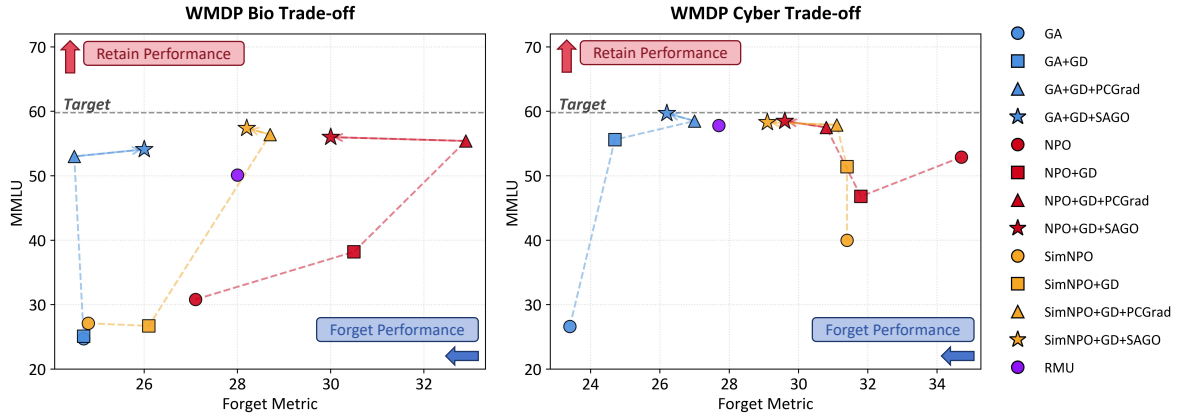

Figure 3: **Performance comparison across different methods on WMDP benchmark.** The plot provides a visualization of the trade-offs between forget and retain performance on WMDP Bio (left) and Cyber (right). A smaller forget metric indicates better forgetting effectiveness, while a larger MMLU value reflects better retention performance. Dashed lines connect base methods to their enhanced variants within the same family (same color). The horizontal grey dashed line represents the original model's performance (Target). The reported results highlight that SAGO (stars) variants consistently push the Pareto frontier upward for comparable forgetting effectiveness.

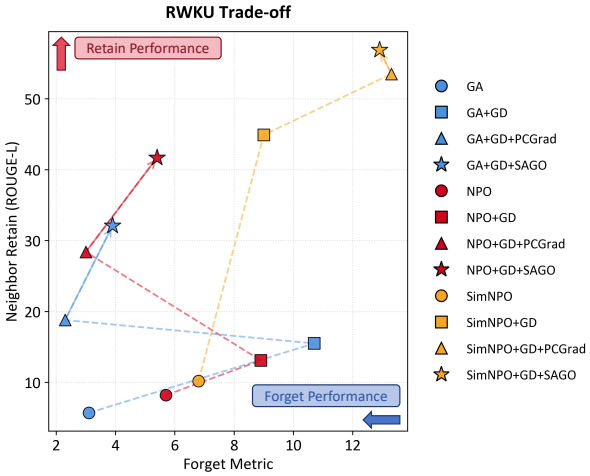

Figure 4: **Performance comparison across different methods on RWKU benchmark.** The plot provides a visualization of the trade-offs between forget and retain performance on RWKU. A smaller forgetting ROUGE-L indicates better forgetting effectiveness, while a larger retention ROUGE-L reflects better retention performance. The reported results highlight that SAGO produces a better frontier than baselines and PCGrad.

Diff (Liu et al., 2022) (equivalent to GA + GD), NPO + GD, and SimNPO + GD. For this category of baselines, we evaluate the effectiveness of PCGrad and SAGO to resolve conflicts between the forget and retain tasks. For WMDP, we also include the RMU (Li et al., 2024) as a baseline, which is proposed alongside the WMDP benchmark.

Evaluation. Following Li et al. (2024), for WMDP, we report forget effectiveness as accuracy on the benchmark's multiple-choice questions, and retention as accuracy on MMLU (Hendrycks et al., 2021). For RWKU, we use fill-in-the-blank (FB) and question-answer (QA) style probes to evaluate the model's ability to recall knowledge and apply it to downstream tasks. Following Jin et al. (2024), we prompt the unlearned model to answer these probes and use ROUGE-L recall score to measure the similarity between the model's predictions and the ground truth answers. The forget performance is evaluated on the 50 forget targets, and the retention is measured on the holdout neighbor targets of these forget targets.

4.2 Experimental Results

We present the full quantitative results in Tables 1, and highlight three observations.

(1) Retention-prioritized gradient synthesis markedly improves the trade-off. Across many settings, integrating a retain objective ("+GD") already improves retention over pure forgetting (GA / NPO / SimNPO). However, retention-prioritized synthesis is decisive: replacing naive summation with PCGrad yields large gains, and SAGO further improves retention while preserving competitive forgetting. For instance, on WMDP Bio with SimNPO+GD, MMLU rises from 26.7 (naive) to 56.4 (+PCGrad) and 57.4 (+SAGO), recovering 96.0% of the target model performance (59.8) while maintaining low forget accuracy (28.2 vs 26.1 baseline). Similar patterns hold in Cyber and RWKU as well.

(2) SAGO consistently matches or exceeds PCGrad on retention with minimal sacrifice in forgetting. On WMDP Bio, SAGO improves MMLU over PCGrad for every base objective: +1.1 (GA+GD), +0.6 (NPO+GD), +1.0 (SimNPO+GD). Cyber likewise exhibits consistent retention gains over PCGrad (+1.2, +1.0, +0.4). Forget performance is also favorable compared to PCGrad, with only an exception on WMDP Bio GA+GD, where the small increase is negligible given the large retention gain. On RWKU, the advantage widens: Neighbor retention (All) improves by +13.3 (GA+GD), +13.3 (NPO+GD), and +3.4 (SimNPO+GD) over PCGrad. These results align with our design goal: retain gradients are never opposed, yielding higher directional fidelity.

(3) Trade-off frontiers shift outward with SAGO. As shown by the Pareto fronts in Figures 3 (WMDP) and Figures 4 (RWKU), SAGO expands the performance envelope: at comparable forgetting levels (*e.g.*, Bio forget ≈ 28), SAGO achieves substantially higher retention. On RWKU, SAGO strictly dominates prior points, defining a new frontier in the retention–forgetting trade-off.

Ablation perspective. The gap between PCGrad and SAGO isolates the contribution of sign-aligned gating beyond orthogonal projection. PCGrad mitigates gradient conflicts through orthogonal projection but may also weaken gradient components that support retention because they are entangled with those conflicts. SAGO preserves these useful directions while ensuring no parameter update opposes the retain gradient direction, empirically yielding superior retention performance.

4.3 Comparison with Other Conflict Mitigation LLM Unlearning Methods

We compare our method with two conflict-mitigation approaches for LLM unlearning: GRU (Wang et al., 2025) and BLUR (Reisizadeh et al., 2025). For each baseline, we report the best-performing variant identified in the respective papers. We also include Global PCGrad, which performs projection in the joint parameter space (as in GRU), to contrast with our module-wise PCGrad that projects per module. As shown in Table 2, our module-wise PCGrad already outperforms Global PCGrad, confirming that finer-grained projection better mitigates inter-task interference. SAGO further improves retention while maintaining competitive forgetting effectiveness.

Method	Forget ↓	MMLU ↑
WMDP Bio		
NPO + GD + GRU	26.2	42.1
RMU + BLUR	27.6	53.5
GA + GD + Global PCGrad	24.8	51.0
GA + GD + PCGrad	**24.5**	53.0
GA + GD + SAGO	26.0	**54.1**
WMDP Cyber		
NPO + GD + GRU	**25.2**	55.6
RMU + BLUR	25.6	55.6
GA + GD + Global PCGrad	28.1	58.1
GA + GD + PCGrad	27.0	58.5
GA + GD + SAGO	26.2	**59.7**

Table 2: **Comparison with other conflict mitigation methods on WMDP.** Lower forget accuracy indicates better forgetting effectiveness, and higher MMLU reflects better retention. PCGrad denotes our module-wise variant.

Method	Forget-Retain	Comb-Forget	Comb-Retain
GradDiff	−0.40	**0.50**	0.42
PCGrad	−0.52	0.29	0.52
SAGO	−0.35	0.17	**0.57**

Table 3: **Average cosine similarity among task and synthesized gradients.** Higher Comb-Retain and moderately positive Comb-Forget are desirable for retention-prioritized unlearning.

4.4 Analysis of Gradient Geometry

To better understand how different synthesis strategies reshape optimization dynamics, we tracked cosine similarities between the raw task gradients (forget and retain) and the final combined gradient ("Comb"). We report the average over 100 training steps on WMDP Cyber in Table 3. **The Forget-Retain similarity** is negative for all methods, which means the two raw tasks naturally pull the model in opposite directions. This conflict is weakest for SAGO, suggesting that SAGO reshapes the parameter space to reduce conflict between forget and retain gradients. **Considering how the final gradient** aligns with the retain gradient (Comb-Retain), SAGO yields the greatest similarity. This matches our goal: never move in a direction that goes against retention. GradDiff attains the largest Comb-Forget similarity, but at the cost of a clear drop in Comb-Retain, meaning the forgetting signal dominates and degrades retention. PCGrad reduces Comb-Forget by projecting away conflicting components. SAGO goes further:

it keeps only gradient components whose signs already agree with the retain task, yielding (i) the strongest alignment with retention (highest Comb-Retain), and (ii) a controlled, non-excessive contribution from forgetting (moderate Comb-Forget).

5 Related Work

LLM Unlearning. The rise of large language models has raised significant concerns about safety risks and privacy leaks, increasing interest in methods for LLM unlearning (Yao et al., 2024). LLM unlearning has a wide range of practical applications, including the protecting copyrighted materials and removing sensitive personal information. A variety of methods for LLM unlearning have been proposed, ranging from model optimization-based (Zhang et al., 2024; Fan et al., 2024; Wang et al., 2025; Sondej et al., 2025) methods to inference-time methods (Pawelczyk et al., 2024; Suriyakumar et al., 2025; Ji et al., 2024). Despite these advances, unlearning in large language models remains challenging because it requires balancing two competing objectives: removing targeted information while preserving the model's overall capabilities (Maini et al., 2024; Shi et al., 2024; Jin et al., 2024; Li et al., 2024). Efforts to address this challenge fall into three categories: (1) objectives that transform the unbounded GA objective into a bounded form to prevent excessive forgetting, uch as NPO (Zhang et al., 2024) and SimNPO (Fan et al., 2024); (2) approaches like GradDiff (Liu et al., 2022) that incorporate an explicit retain objective; and (3) inference-time unlearning methods that modify model outputs without altering model weights (Pawelczyk et al., 2024; Suriyakumar et al., 2025; Ji et al., 2024). Currently, GRU (Wang et al., 2025) uses PCGrad (Yu et al., 2020) to resolve gradient conflicts between forget and retain gradients. Our work builds on similar ideas but focuses on the asymmetric nature of the two tasks.

Conflict Mitigation in Multi-task Learning. Multi-task learning (MTL) refers to learning a single model that can tackle multiple different tasks. However, learning multiple tasks simultaneously can be a challenging optimization problem. The most common MTL objective in practice is the weighted loss over all tasks. But directly optimizing the weighted loss is known to lead to undesirable performance. A known cause of this phenomenon is the conflicting gradients between different tasks (Yu et al., 2020). To address this problem, previous works have proposed various gradient manipulation techniques (Yu et al., 2020; Liu et al., 2021; Chen et al., 2020; Navon et al., 2022). For example, PCGrad (Yu et al., 2020) seeks a better update vector by projecting one task's gradient onto the normal plane of another task's gradient. Another line of work is multi-task learning via model merging (Yang et al., 2024). Instead of manipulating gradients, these methods first train separate models for each task and then merge the models into a single one. To mitigate the conflict between different tasks, a variety of methods have been proposed to resolve conflicts among task vectors (Yadav et al., 2023; Gargiulo et al., 2025; Yu et al., 2024; Marczak et al., 2025; Sun et al., 2025).

6 Conclusion

In this paper, we revisited LLM unlearning through an *asymmetric two-task* perspective. Building on this perspective, we introduced a modular framework for conflict-aware gradient synthesis and instantiated it with (i) a per-parameter adaptation of PCGrad and (ii) SAGO, a novel gradient synthesis method that gates forget and retain signals at element level to guarantee non-antagonistic updates. Across WMDP and RWKU, SAGO consistently shifts the Pareto frontier outward, recovering substantially more retention at comparable forgetting strength, validating that retention-prioritized gradient geometry matters.

Limitations

Our study, while demonstrating consistent retention gains under many settings, still has several limitations. (1) Benchmark scope: We focus on two representative LLM unlearning benchmarks (WMDP and RWKU); broader domains (*e.g.*, multimodal or code models) are left for future exploration. (2) Computational cost: Computing and storing separate retain and forget gradients introduces additional overhead versus a single objective, though still practical in standard finetuning regimes; we do not optimize for extremely low-resource settings.

Acknowledgements

This project was supported by National Natural Science Foundation of China (No. 62306132), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011564), Natural Science Foundation of Shanghai (No. 25ZR1402136). We thank the anonymous reviewers for their insightful

feedback on this work. This work was done during Zeguan's internship at SUSTech.

References

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. 2020. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *Advances in Neural Information Processing Systems*, 33:2039–2050.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2024. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. 2025. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

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

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

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Preprint*, arXiv:2406.10890.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, and 1 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

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

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. 2021. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890.

Xiaoqian Liu, Yangfan Du, Jianjin Wang, Yuan Ge, Chen Xu, Tong Xiao, Guocheng Chen, and Jingbo Zhu. 2025. A modular-based strategy for mitigating gradient conflicts in simultaneous speech translation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost van de Weijer. 2025. No task left behind: Isotropic model merging with common and task-specific subspaces. *arXiv preprint arXiv:2502.04959*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. 2022. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16428–16446. PMLR.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, pages 40034–40050.

Hadi Reisizadeh, Jinghan Jia, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, Sijia Liu, and Mingyi Hong. 2025. Blur: A bi-level optimization approach for llm unlearning. *arXiv preprint arXiv:2506.08164*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way

evaluation for language models. *arXiv preprint arXiv:2407.06460*.

Filip Sondej, Yushi Yang, MikoŁ Kniejski, Marcel Windys, and 1 others. 2025. Robust llm unlearning with mudman: Meta-unlearning with disruption masking and normalization. *arXiv preprint arXiv:2506.12484*.

Wenju Sun, Qingyong Li, Yangli-ao Geng, and Boyang Li. 2025. Cat merging: A training-free approach for resolving conflicts in model merging. *arXiv preprint arXiv:2505.06977*.

Vinith M Suriyakumar, Ayush Sekhari, and Ashia Wilson. 2025. Ucd: Unlearning in llms via contrastive decoding. *arXiv preprint arXiv:2506.12097*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. 2025. GRU: Mitigating the trade-off between unlearning and retention for LLMs. In *Forty-second International Conference on Machine Learning*.

Zeguan Xiao, Yan Yang, Guanhua Chen, and Yun Chen. 2024. Distract large language models for automatic jailbreak attack. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16230–16244, Miami, Florida, USA. Association for Computational Linguistics.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Yan Yang, Zeguan Xiao, Xin Lu, Hongru Wang, Xuetao Wei, Hailiang Huang, Guanhua Chen, and Yun Chen. 2025. SeqAR: Jailbreak LLMs with sequential auto-generated characters. 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)*, pages 912–931, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *Preprint*, arXiv:2308.06463.

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

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Hyperparameters

Since the evaluation of unlearning involves two opposing metrics—forgetting and retention—there is an inherent trade-off between them. Our hyperparameter tuning protocol is therefore to first match the forgetting metrics across methods as closely as possible, and then compare the methods under this constraint. Concretely, we mainly tune the following hyperparameters:

- Weight γ of the forgetting objective: the default value is 1.0, and we sweep $\gamma \in [0.1, 1.0]$ when necessary to match the baseline's forgetting performance.
- Weight α of the retention objective: the default value is 1.0, and we sweep $\alpha \in [0.1, 1.0]$ when necessary to match the baseline's forgetting performance.
- Training budget and learning rate: for WMDP, the default is 100 training steps; for RWKU, we fix the budget to 2 epochs and tune the learning rate and method-specific hyperparameters (*e.g.*, β for NPO).