Leveraging Machine Unlearning for Cost-Efficient Preference Alignment

Xiaohua Feng[1], Yuyuan Li[2], Huwei Ji[1], Li Zhang[1], Jiaming Zhang[1], Tianyu Du[1], Chaochao Chen[1,*]

Zhejiang University[1] Hangzhou Dianzi University[2]

Introduction

Background

Preference Alignment (PA) guides LLMs to generate responses aligned with human preferences, such as helpfulness, harmlessness, and factuality. However, mainstream RLHF-style methods require costly positive preference examples and expensive policy optimization, making them difficult to apply in low-resource alignment scenarios.

Key Opportunity

Instead of learning from many positive examples, Machine Unlearning (MU) offers a cheaper alternative: directly remove the influence of undesirable negative examples.

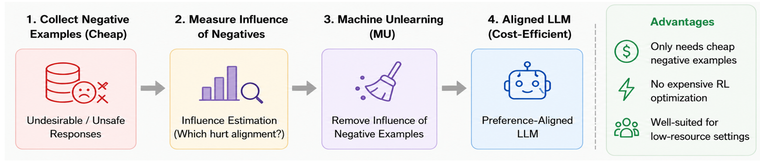

Figure 1. MU provides a cost-efficient route to PA.

Takeaway. The central question is not only whether negative examples can be unlearned, but which ones should be unlearned to best improve PA.

Preference Alignment Objective

Given policy π_θ, PA seeks high expected reward:

$$J(\theta) = \mathbb{E}_{\pi_\theta}[r(x_{<t}, x_t)]\,. \tag{1}$$

LLM Unlearning Objective

Given forget set $\mathcal{D}_f$, LLM unlearning can be written as:

$$\min_\theta \quad \underbrace{L_F(\mathcal{D}_f;\theta)}_{\text{forget negative data}} + \lambda \quad \underbrace{L_R(\theta)}_{\text{preserve utility}} \quad . \tag{2}$$

From MU to PA: A Bi-level View

Sample-Level Unlearning Impact

For a negative sample x with unlearning weight ω:

$$\theta^*(\omega) = \arg\min_\theta \omega L_F(x;\theta) + \lambda L_R(\theta). \tag{3}$$

The corresponding PA change is approximated by:

$$\Delta J(\theta^*(\omega)) \approx -\frac{\omega}{2}\nabla_\theta J(\theta^*)^\top \nabla_\theta L_F(x;\theta^*). \tag{4}$$

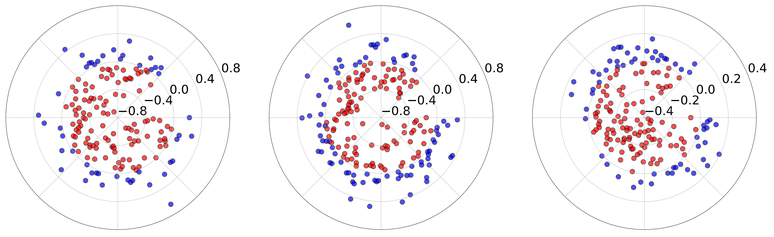

Figure 2. Effect of unlearning individual negative samples on PA performance of the LLaMA2 model.

Observation. Unlearning negative examples does not always improve PA. Different samples can have positive, negative, or weak impact, and the magnitude depends on the unlearning weight.

Method: Unlearning to Align

Core Idea

Do not unlearn all negative examples equally. U2A selects the negative examples that are most beneficial for PA and assigns them optimized unlearning weights.

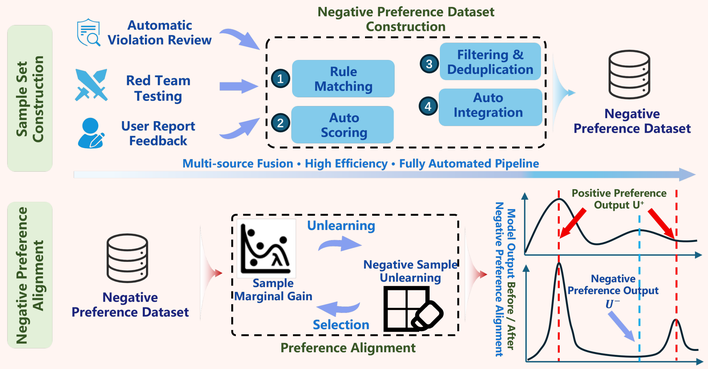

Figure 3. Pipeline of the U2A framework.

Takeaway. U2A transforms MU-based alignment into an unlearning sample selection and reweighting problem.

U2A Optimization Formulation

Given negative samples $\mathcal{D} = \{x_i\}_{i=1}^n$ and sample losses $\ell_i(\theta)$, U2A solves:

$$\min_{\omega\in\Delta_n} -J(\theta^*(\omega)) + \beta L_p(\omega), \tag{5}$$

$$\text{s.t.} \quad \theta^*(\omega) = \arg\min_\theta \sum_{i=1}^n \omega_i \ell_i(\theta) + \lambda L_R(\theta). \tag{6}$$

Sparse reweighting. The sparsity-inducing regularizer is:

$$L_p(\omega) = \sum_{i=1}^n \sqrt{\omega_i}, \quad \omega_i \geq 0, \quad \sum_i \omega_i = 1. \tag{7}$$

- $\omega_i = 0$: sample x_i is not selected for unlearning.
- Larger ω_i: sample x_i contributes more to weighted MU.
- Sparse ω avoids unnecessary unlearning and improves PA.

Efficient Optimization

Step 1: Marginal-Gain Selection

At each iteration, U2A selects samples with the largest outer-objective gain:

$$\Delta g(k) = -\nabla_\theta J(\theta^*)^\top \left(\frac{\partial^2 f}{\partial \theta^2}\right)^{-1} \nabla_\theta \ell_k(\theta^*) - \frac{\beta}{2}\omega_k^{-\frac{1}{2}}. \tag{8}$$

Step 2: Weight Refinement

After fixing the support set S, weights are updated on the simplex by mirror descent:

$$\omega^{t+1} = \arg\min_{\omega\in\Delta_{|S|-1}} \langle \nabla g(\omega^t), \omega\rangle + \frac{1}{\eta} D_h(\omega\|\omega^t). \tag{9}$$

Using negative entropy gives:

$$\omega_i^{t+1} = \frac{\omega_i^t \exp(-\eta\nabla_i g(\omega^t))}{\sum_{j\in S}\omega_j^t \exp(-\eta\nabla_j g(\omega^t))}, \quad i \in S. \tag{10}$$

Experiments

Setup

- **Tasks:** reducing harmfulness, enhancing usefulness, eliminating hallucinations.
- **Datasets:** SafeRLHF, UltraFeedback, HaluEval.
- **Models:** LLaMA2-7B-Chat, LLaMA3.1-8B-Instruct, Qwen2.5-14B.
- **Baselines:** Retrain, GA, GradDiff, NPO, PPO, DPO.

SafeRLHF: U2A Improves MU Baselines

Method	Reward-V↑	ASR-A↓	ASR-S↓
	LLaMA3		
Original	-19.827	0.848	0.794
Retrain	-17.911	0.740	0.686
GA	-18.011	0.829	0.750
GA + U2A	**-7.609**	0.162	0.121
GradDiff	-18.477	0.829	0.767
GradDiff + U2A	-12.644	**0.103**	0.173
NPO	-17.985	0.860	0.785
NPO + U2A	-13.815	0.698	0.498

Takeaway. U2A substantially improves PA reward and reduces ASRs.

UltraFeedback: Competitive with PA Methods

Method	Reward-V↑	LC-WR↑	GPT4-WR↑	Coh.↑
	LLaMA3			
Original	-4.996	10.57	5.85	0.726
PPO	-1.302	23.43	19.41	0.740
DPO	-0.277	**25.30**	19.08	0.738
GA + U2A	1.105	24.79	**21.33**	**0.745**
GradDiff + U2A	**1.248**	24.18	19.67	0.738
NPO + U2A	-0.248	20.56	13.20	0.729

Takeaway. U2A can match or outperform PPO/DPO on several PA metrics.

Ablation and Sensitivity

High-Gain Samples Matter

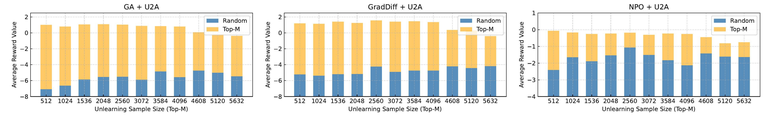

Figure 4. Top-K marginal-gain selection outperforms random selection.

Finding. Marginal-gain-based Top-K selection consistently outperforms random selection. Adding too many low-gain samples leads to saturation or degradation.

Overall Conclusion

- MU is a viable route to PA when positive preference data is limited.
- Negative examples have heterogeneous and sometimes adverse effects.
- U2A improves PA by learning **what & how to unlearn** and **how strongly to unlearn it**.