Elastic Robust Unlearning of Specific Knowledge in Large Language Models

Yize Sui, Jing Ren, Wenjing Yang, Ruochun Jin, Liyang Xu, Xiyao Liu, Ji Wang [†]

The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)

Contents

1. Background

Why do we need "LLM Unlearning"?

- The Dilemma of LLMS
  - The massive pre-training data inevitably contains harmful, infringing and privacy content.
- The shortcomings of traditional solutions
  - Safety Retraining: extremely costly and unrealistic.
  - fine-tuning: It is only effective in surface behavior and fails to remove information from the model's knowledge level.

LLM Unlearning: The goal is to directly and efficiently remove specific knowledge from model parameters.

1. Background

- Core challenge:
  - How can one forget things completely?
  - How can one not forget what should not be forgotten?
  - How to prevent being "awakened"?

2. Motivation

Existing work and our motivation

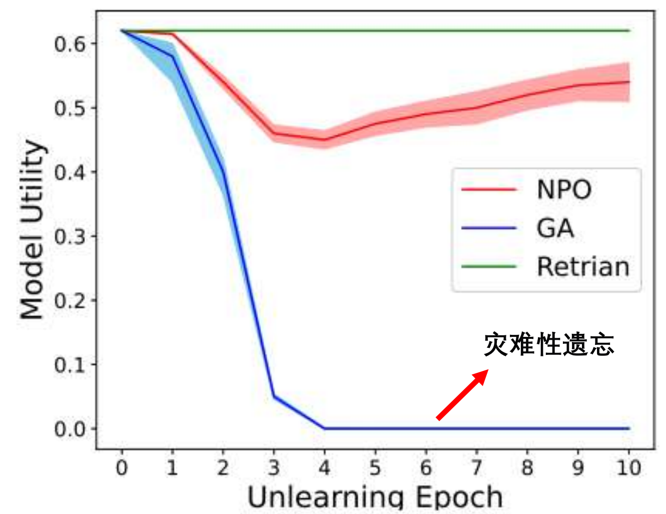
- Method evolution
  - Gradient Ascent (GA) ： It can easily lead to model collapse
  - Negative Preference Optimization (NPO)：Consider unlearning as a special type of "preference" learning.

$$\mathcal{L}_{\mathrm{GA}}(\pi_\theta) = -\underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{f}}}\left[-\log\left(\pi_\theta(y \mid x)\right)\right]}_{\text{prediction loss}},$$

unlearning模型

$$\mathcal{L}_{NPO}\left(\pi_\theta, \pi_{\mathrm{ref}}\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{f}}}\left[-\frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\boxed{\pi_\theta}(y \mid x)}{\boxed{\pi_{\mathrm{ref}}}(y \mid x)}\right)\right]$$

参考模型



(b) Model utility of NPO and GA across epochs.

2. Motivation

Existing work and our motivation

- Two core flaws of PO-based unlearning:
  - Rigid reward setting:
    - **reference-based reward**: In the early stage of training, the smoothing of gradient weights fails, behaving like GA and damaging utility.
    - **reference-free reward**: Using a constant offset, a uniform distribution, instead of the reference model, the specific differences between samples are lost.

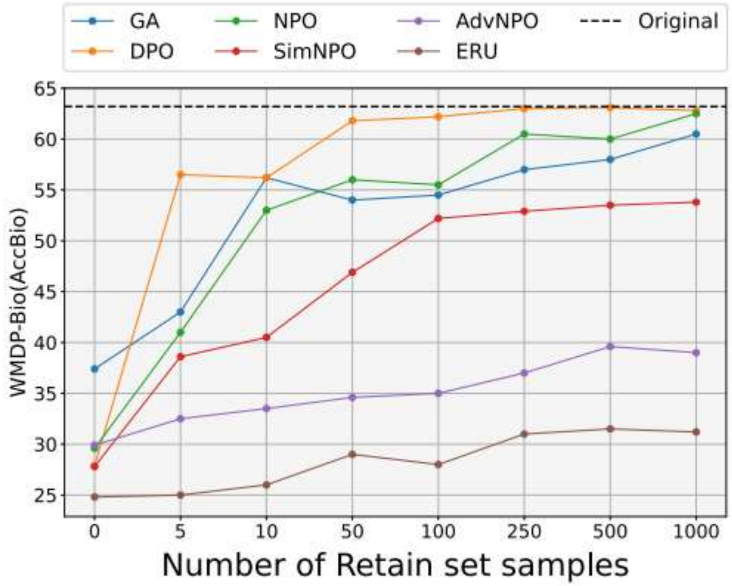
$$\ell_{\mathrm{NPO}}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{f}}} \underbrace{\left[-\frac{2}{\beta} \log \sigma \left(-\beta \log \left(\frac{\pi_{\boldsymbol{\theta}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} \right) \right) \right]}_{① := \ell_{\mathrm{f}}(y|x;\boldsymbol{\theta}),\ \text{the specified forget loss in (1)}},$$

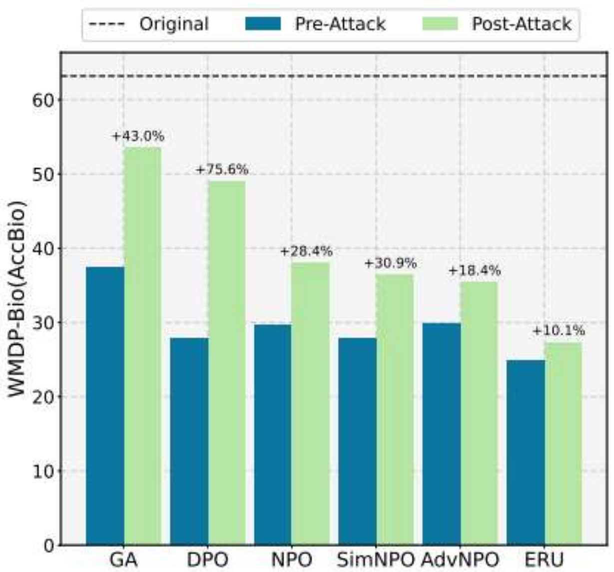
$$\ell_{\mathrm{SimNPO}}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{f}}} \left[-\frac{2}{\beta} \log \sigma \left(-\frac{\beta}{|y|} \log \pi_{\boldsymbol{\theta}}(y|x) - \gamma \right) \right],$$

2. Motivation

Existing work and our motivation

- Two core flaws of PO-based unlearning:
  - Lack of unlearning robustness
    - ➢ Relearning attack: Fine-tuning with just 10 irrelevant samples can significantly restore forgotten knowledge.
    - ➢ Adversarial attack: By carefully designing prompt words, the forgetting limit can be bypassed.

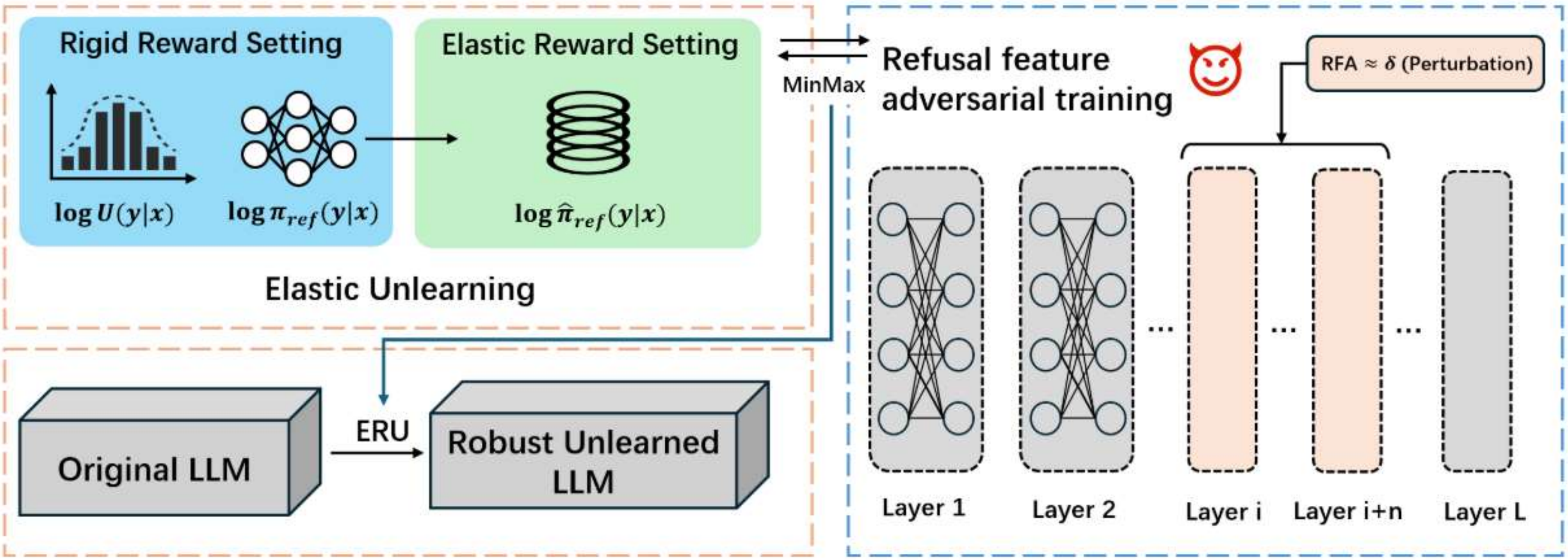




3. Method

Elastic Robust Unlearning (ERU)

- Two core pillars
  - Elatsic reward setting: Balance the two reward signals
  - Refusal Feature Adversarial Training: Simulating the worst-case scenario of knowledge recovery during training to enhance unlearning robustness



3. Method

Elastic Robust Unlearning (ERU)

- Elatsic Reward Setting
  - By combining the advantages of the reference-based reward and reference-free reward, the reward weights are dynamically adjusted.
  - Substitute the NPO loss function and perform length normalization to obtain the new objective function:

$$\hat{\pi}_{\mathrm{ref}}(y|x) = U(y|x)\left(\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)^{\alpha}$$

$$\mathcal{L}_{NPO}\left(\pi_\theta, \pi_{\mathrm{ref}}\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{f}}}\left[-\frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta\left(y \mid x\right)}{\boxed{\pi_{\mathrm{ref}}\left(y \mid x\right)}}\right)\right]$$

3. Method

Elastic Robust Unlearning (ERU)

- Elatsic Reward Setting
  - Substitute the NPO loss function and perform length normalization to obtain the new objective function:

$$
\begin{aligned}
&\mathcal{L}^{new}\left(\pi_\theta, \hat{\pi}_{\text{ref}}, U\right) \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[-\frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(y\mid x)}{\hat{\pi}_{\text{ref}}(y\mid x)}\right)\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[-\frac{2}{\beta}\log\sigma\left(-\beta\log\pi_\theta(y\mid x) - M\right)\right]
\end{aligned}
$$

$$
M = \left[\gamma' - \alpha\left(\frac{\beta\log\left(\frac{\pi_\theta(y\mid x)}{\pi_{\text{ref}}(y\mid x)}\right) - \mu^*}{\sigma^*}\right)\right]
$$

$$
\gamma' = \beta\left(-\log U(y\mid x)\right)
$$

length normalization

$$
\mathcal{L}_{EU}\left(\pi_\theta, \hat{\pi}_{\text{ref}}, U\right) = \mathbb{E}_{(x,y)\in\mathcal{D}_f}\left[-\frac{2}{\beta}\log\sigma\left(u(x,y) - \text{rg}\left[M\right]\right)\right]
$$

$$
u(x,y) = -\frac{\beta}{|y|}\log\pi_\theta(y \mid x)
$$

$$
W'_\theta(x,y) = \left(\frac{2\cdot exp(\text{rg}\left[M\right])\cdot\left[\pi_\theta(y\mid x)\right]^{\frac{\beta}{|y|}}}{exp(\text{rg}\left[M\right])\cdot\left[\pi_\theta(y\mid x)\right]^{\frac{\beta}{|y|}} + 1}\right)\cdot\boxed{\frac{1}{|y|}}
$$

3. Method

Elastic Robust Unlearning (ERU)

- Refusal Feature Adversarial Training (RFAT)
  - Refusal Feature
    - Research has found that the ability of LLMS to identify harmful problems largely depends on a specific and locatable direction vector (refusal feature) in the activation space of their hidden layers.
    - Arditi et al. demonstrated that the key mechanism of adversarial perturbation is to eliminate the refusal feature. (refusal feature ablation)。
    - Based on these findings, RFAT effectively conducts LLM adversarial training by simulating the effect of adversarial attacks through RFA.

3. Method

Elastic Robust Unlearning (ERU)

- Refusal Feature Adversarial Training
  - Refusal Feature Ablation
    - Basic idea: The r refusal feature is captured by comparing the internal activation differences when the model processes harmful and harmless inputs.

$$\mathbf{r}_{\mathrm{HH}}^{(l)} = \frac{1}{|\mathcal{D}_{\mathrm{harmful}}|} \sum_{x \in \mathcal{D}_{\mathrm{hamful}}} \mathbf{h}^{(l)}(x) - \frac{1}{|\mathcal{D}_{\mathrm{harmless}}|} \sum_{x \in \mathcal{D}_{\mathrm{harmless}}} \mathbf{h}^{(l)}(x)$$

3. Method

Elastic Robust Unlearning (ERU)

- Refusal Feature Adversarial Training
  - RFAT applied in LLM unlearning
    - We describe the adversarial training applied to LLM unlearning as a minimax optimization problem.

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{f}}} \max_{\delta} \mathcal{L}\left(\pi_{\theta}\left(x+\delta, y\right)\right)$$

通过RFA模拟

Thanks!

Yize Sui, Jing Ren, Wenjing Yang, Ruochun Jin, Liyang Xu, Xiyao Liu, Ji Wang †