A Robust Unlearning Method with Adaptive Knowledge Guidance and Memory Preservation

Jingyuan Tian[1,2], **Xiaofei Zhou**[1,2*]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
{tianjingyuan, zhouxiaofei}@iie.ac.cn

Abstract

Machine unlearning has emerged as a promising approach to remove specific knowledge from large language models (LLMs), especially for safety-critical applications. However, existing representation-based methods lack guidance for selecting representation locations to unlearn (RMU), while probability-based methods are vulnerable to fine-tuning attacks which use unrelated and safe data to fine-tune models. To address these problems, this paper presents an Adaptive Localized Memory Perturbation Unlearning (ALMPU) method, which uses knowledge guidance and adversarial memory perturbation to improve unlearning robustness. Specifically, we apply scaling factors to attention heads and select the most sensitive ones as knowledge guidance. Guided by knowledge localization, we integrate enhanced memory perturbation into the standard representation-based unlearning process to preserve specific knowledge. By adding interventions to selected attention heads and explicitly optimizing against fine-tuning attacks during the unlearning process, ALMPU creates a controlled divergence from the original model that is inherently resistant to relearning attempts. Experimental evaluation on the WMDP benchmark demonstrates that ALMPU consistently outperforms baseline methods across different scales of fine-tuning attacks.

Introduction

Large language models (LLMs) have achieved remarkable capabilities but inevitably retain harmful or sensitive information from their vast training data(Maini et al. 2024; Jin et al. 2024). This poses significant risks in safety-critical applications(Yao et al. 2024), where generation of dangerous content could have serious real-world consequences. Machine unlearning(Cao and Yang 2015; Yao, Xu, and Liu 2023) has emerged as a promising solution to selectively remove specific knowledge while preserving general capabilities(Liu et al. 2024), going beyond simple output filtering to modify internal model representations(Liu et al. 2025).

Recent research demonstrates that existing unlearned models can have their unlearning effects compromised by adversaries using very small samples of unrelated datasets (Łucki et al. 2024; Barez et al. 2025), as illustrated in Figure 1. Left two figures (Li et al. 2024) illustrate that unlearned model refuses to generate harmful content because unlearning method is applied to remove harmful knowledge. Right figure reveals that the unlearning effect can be compromised through fine-tuning on small amounts of safe data. This problem stems from the inherent limitations of existing unlearning methods. Current methods can be categorized into representation-based and probability-based methods (Yuan et al. 2024; Shumailov et al. 2024a).

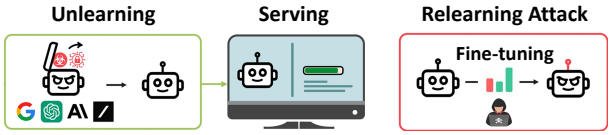

Figure 1: Overview of the fine-tuning attack vulnerability in unlearned models

Representation-based Methods use intermediate layer feature representations of the model for unlearning. Representation Misdirection for Unlearning (RMU) (Li et al. 2024) manipulates model activations at selected layers rather than modifying all parameters, employing a joint objective that manipulates activation norms of harmful data to a random vector to remove them from the model while preserving them on benign data to maintain the utility of target model. RMU intuitively selects model representations in early layers without theoretical foundation (Yin, Ye, and Durrett 2024), and parameters from a single layer are insufficient to cover vulnerable areas across multiple layers, making the model vulnerable to fine-tuning (Jia et al. 2024).

Probability-based Methods use the model's probability of outputting harmful content as guidance for optimization (Suriyakumar, Sekhari, and Wilson 2025; Ding et al. 2025). Gradient Ascent (GA) directly optimizes against harmful content generation by ascending gradients with respect to the forget set. Negative Preference Optimization (NPO) (Zhang et al. 2024) treats unlearning as a preference learning problem, training models to prefer non-harmful over harmful responses. Direct Preference Optimization (DPO) (Rafailov et al. 2024) and its variants align models to generate specific safe responses when queried about harmful content (Feng et al. 2025). While these methods can achieve effective knowledge removal, they often suffer from fine-

*Corresponding author.

tuning attacks (Łucki et al. 2024; Yuan et al. 2025), as they need to balance model utility while unlearning across the entire model without providing explicit safety protection in vulnerable areas. This leads to rapid degradation of the model's unlearning performance when attacked, as even small amounts of data can exploit vulnerable areas to quickly compromise unlearning (Gao et al. 2023).

In this paper, we address this challenge by introducing ALMPU (Adaptive Localized Memory Perturbation Unlearning). Current representation-based methods apply interventions to random layers while being vulnerable to fine-tuning attacks that target the entire network while probability-based methods are vulnerable to fine-tuning attacks. To adaptively find the vulnerable area, we identify the most sensitive attention heads for harmful knowledge processing through an adaptive sensitive attention head selection mechanism that learns element-wise scaling factors and select those with highest paramter difference. These are the top-K attention heads most relevant to harmful content generation (Yin, Ye, and Durrett 2024). They provide knowledge guidance of which components are most responsible for harmful outputs. To enhance the robustness of unlearned model against fine-tuning attacks, we use adversarial memory perturbation training strategy. Based on knowledge guidance, ALMPU incorporates memory perturbation as a defense mechanism at vulnerable locations. This perturbation enables the model to shift from the original single rejection direction to searching for robust unlearning directions within a perturbation-radius norm ball during the standard forgetting process, making the model more resistant to attacks from fine-tuning with small amounts of unrelated samples. Experimental results on WMDP benchmark and Wikitext datasets demonstrate that our method achieves effective unlearning while maintaining robustness against fine-tuning attacks. Our key contributions are presented as follows:

- We propose an adaptive knowledge guidance mechanism that searches sensitive attention head which learns element-wise scaling factors to identify the top-K attention heads most critical for harmful knowledge processing.
- We propose a memory perturbation training strategy at vulnerable locations that enables the model to search for robust unlearning directions within a perturbation-radius norm ball, shifting from single rejection directions to more resilient unlearning patterns that resist fine-tuning attacks with unrelated data.
- We demonstrate that ALMPU maintains unlearning effectiveness after various size of adversarial fine-tuning attacks on the WMDP benchmark, outperforming baseline methods by providing explicit safety protection across vulnerable areas spanning multiple layers.

Related Work

Machine unlearning methods for LLMs can be categorized into gradient-based and representation-based approaches based on their underlying mechanisms (Yuan et al. 2024).

Gradient-based Methods Gradient-based methods modify full model parameters through loss optimization on forget and retain sets. Gradient Ascent (GA) represents the most fundamental approach, performing gradient ascent on the forget set to reverse the original training process, essentially "undoing" the learning of harmful knowledge. However, GA often leads to unstable training and catastrophic forgetting due to its aggressive parameter updates.

Negative Preference Optimization (NPO) introduces a more stable alternative by treating unlearning as a preference learning problem, training models to "anti-prefer" harmful outputs rather than directly maximizing loss on them:

$$\mathcal{L}_{NPO}(\mathcal{D}_F;\theta) = -\frac{2}{\beta}E_{(x,y)}\left[\log\sigma\left(-\beta\frac{p(y|x;\theta)}{p(y|x;\theta_{ref})}\right)\right]$$

where β controls the deviation from the reference model θ_{ref} and $(x, y) \sim \mathcal{D}_F$. The sigmoid saturation provides controlled logarithmic divergence.

Direct Preference Optimization (DPO) adapts the standard preference optimization framework for unlearning by treating forget set answers as negative samples and rejection templates as positive samples, enabling models to learn explicit refusal patterns for harmful queries. However, gradient-based methods require extremely small learning rates during fine-tuning to reduce parameter update magnitude and preserve model utility, making the unlearning effect vulnerable to fine-tuning attacks.

Representation-based Methods Representation-based methods target specific model components rather than modifying all parameters. Representation Misdirection for Unlearning (RMU) manipulates model activations at early layers (layer 7 for zephry-7B model). This work proposes that activation norms in earlier layers make it difficult for later layers to process hazardous information effectively:

$$\mathrm{L}_{RMU} = E_{x_f}\left[\frac{1}{L_f}\sum_{t\in x_f}\|M_{updated}(t) - c\cdot u\|_2^2\right] + \alpha E_{x_r}\left[\frac{1}{L_r}\sum_{t\in x_r}\|M_{updated}(t) - M_{frozen}(t)\|_2^2\right]$$

where u is a random unit vector, c controls activation scaling, and α balances the forget and retain objectives, $x_f \sim D_{forget}$, $x_r \sim D_{retain}$. This method maintains the balance between unlearning effect and model utility which is a common trade-off (Zhong, Luo, and Liu 2025). However, RMU's layer selection strategy is result-oriented and lacks theoretical guidance.

WMDP Benchmark for Hazardous Knowledge Evaluation Unlike other datasets mainly focusing on privacy(Maini et al. 2024; Jin et al. 2024; Shi et al. 2024), the Weapons of Mass Destruction Proxy (WMDP) benchmark (Li et al. 2024) provides the first comprehensive public framework for evaluating hazardous knowledge in large language models. WMDP consists of 3,668 expert-written multiple-choice questions across biosecurity, cybersecurity, and chemical security domains, developed through collaboration between academics and technical consultants. The benchmark underwent stringent filtering to remove sensitive information while maintaining utility as a proxy measurement for hazardous capabilities. WMDP serves dual purposes: evaluating dangerous knowledge presence in models and providing a standardized testbed for unlearning methods, establishing it as the standard evaluation tool for both

hazardous capabilities assessment and safety intervention effectiveness.

Fine-tuning Attacks on Unlearned Models A critical vulnerability of existing unlearning methods is their susceptibility to fine-tuning attacks, where adversaries use additional training data to recover supposedly unlearned knowledge (Qi et al. 2023; Lermen, Rogers-Smith, and Ladish 2023). Research has demonstrated that safety alignment and unlearning effects can be compromised through fine-tuning on seemingly benign datasets, with LoRA-based fine-tuning requiring as few as dozens of examples from unrelated datasets to substantially recover unlearned capabilities. This vulnerability stems from the gradient-based nature of both unlearning and recovery processes, where subsequent gradient updates during fine-tuning can naturally reverse the parameter changes made during unlearning. The efficiency of these attacks makes them particularly concerning, as adversaries can potentially reverse unlearning without access to the original forget set or specialized attack data (Vasilev et al. 2025). Recent work has emphasized the importance of evaluating unlearning methods against fine-tuning attacks as a standard robustness benchmark (Shumailov et al. 2024b), highlighting the need for methods that can maintain robustness against parameter updates while preserving model utility. Our work directly addresses this challenge by incorporating explicit defenses against fine-tuning attacks during the unlearning process.

Method

In this section, we present our ALMPU (Adaptive Localized Memory Perturbation Unlearning) approach for robust machine unlearning, which integrates an adaptive sensitive attention head selection mechanism with enhanced memory perturbation. Our method addresses two key limitations of existing methods: the lack of knowledge guidance in representation-based methods for selecting intervention locations, and the vulnerability to fine-tuning attacks that can rapidly compromise unlearning effectiveness. ALMPU presents a two-phase framework as shown in Figure 2: (1) adaptive attention head selection to identify the most sensitive components for harmful knowledge processing, and (2) alternating memory perturbation training that creates robust resistance against relearning attempts. We begin by detailing each component of our approach.

Localized Attention Head Selection

The motivation for attention head selection stems from recent advances in knowledge editing research, which has revealed that large language models process certain knowledge through specialized pathways (Yin, Ye, and Durrett 2024). This specialization can be transferred to sensitive or harmful content unlearning, where specific attention heads play important roles in content generation and reasoning.

The key insight underlying our approach is that harmful knowledge is not uniformly distributed across all model components but rather concentrated in specific attention heads that serve as bottlenecks for harmful content processing. Formally, let $\mathcal{H}$ denote the set of all attention heads in the model, and let $f_h(x)$ represent the activation magnitude of head h when processing input x. For harmful knowledge $x_h \in \mathcal{D}_{forget}$, we hypothesize that there exists a subset $\mathcal{H}^* \subset \mathcal{H}$ such that:

$$\sum_{h \in \mathcal{H}^*} f_h(x_h) \gg \sum_{h \in \mathcal{H} \setminus \mathcal{H}^*} f_h(x_h) \tag{1}$$

This hypothesis is supported by knowledge editing research showing that specific capabilities are localized to particular model components. Specifically, we introduce learnable scaling factors $A_i^l \in R^{d_{head}}$ for each attention head i at layer l, where d_{head} is the attention head dimension. These scaling factors act as element-wise multipliers that can amplify specific dimensions of the attention head outputs. During the forward pass, the activation $z_t^{(l,i)}$ of attention head i at layer l is modified as:

$$z_t^{(l,i)} \leftarrow (1 + A_i^l) \odot z_t^{(l,i)} \tag{2}$$

where $\odot$ denotes element-wise multiplication. We freeze all pre-trained weights and learn the scaling factors A_i^l end-to-end using the cross-entropy loss on $\mathcal{D}_{forget}$. The scaling factors are initialized from $\mathcal{N}(0, \sigma_A)$ and regularized with an L1 penalty term with coefficient λ to encourage sparsity:

$$\mathcal{L}_{selection} = \mathcal{L}_{CE} + \lambda \sum_{l,i} \|A_i^l\|_1 \tag{3}$$

After training, we score each attention head using the L2 norm of its learned scaling factors: $S(l,i) = \|A_i^l\|_2$. A larger score indicates that the location is more important to process harmful knowledge. We select the top-K attention heads with the highest scores to form our target set $\mathcal{T} = \{(l_1, i_1), \ldots, (l_K, i_K)\}$.

Memory Perturbation for Robustness

While the localized attention head selection identifies vulnerable components, we introduce enhanced memory perturbation as a defense mechanism to strengthen the model's robustness against fine-tuning attacks.

Given a pre-trained language model parameterized by θ, machine unlearning aims to remove specific knowledge $\mathcal{D}_{forget}$ while retaining the model's performance on general knowledge $\mathcal{D}_{retain}$. Each example consists of input-output pairs (x, y) where x represents the input sample of harmful data and y represents model output.

Adversarial Memory Preservation We design adversarial perturbation parameters specifically for the selected sensitive attention heads. For each identified sensitive head (l, h) at layer l and head position h, we initialize learnable perturbation parameters:

$$\delta_{l,h} \sim \mathcal{N}(0, \sigma_{adv}^2 \mathbf{I}) \tag{4}$$

where $\delta_{l,h} \in R^{d_{head}}$ represents the adversarial perturbation for head (l, h), and σ_{adv} controls the initial perturbation magnitude.

During the forward pass, we apply these perturbations to the attention head outputs:

$$\tilde{\mathbf{h}}_{l,h} = \mathbf{h}_{l,h} + \delta_{l,h} \tag{5}$$

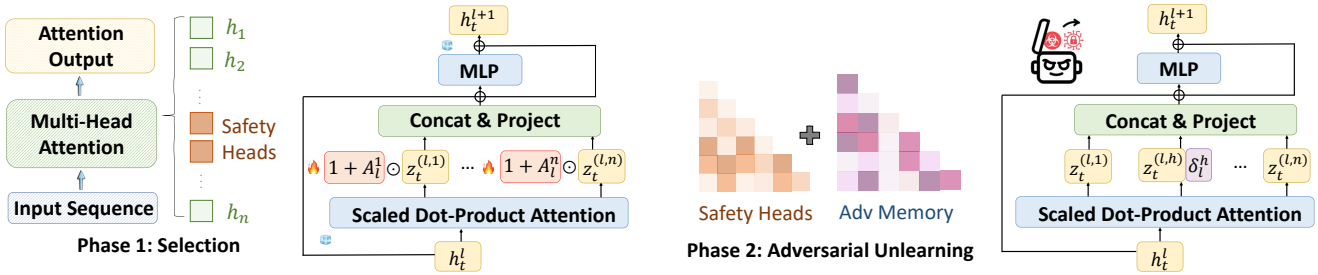

Figure 2: Framework of ALMPU (Adaptive Localized Memory Perturbation Unlearning). Phase 1 demonstrates the adaptive attention head selection mechanism using learnable scaling factors to identify the top-K most sensitive heads for harmful knowledge processing. Phase 2 illustrates the memory perturbation training that applies adversarial perturbations to selected heads while unlearning.

where $\mathbf{h}_{l,h}$ is the original attention head output and $\tilde{\mathbf{h}}_{l,h}$ is the perturbed output.

Multi-Stage Optimization Strategy Our training process employs a two-stage optimization strategy that separately optimizes model parameters and adversarial perturbations to achieve both effective unlearning and robust defense.

Stage 1: Layer-level Optimization In the first stage, we optimize the standard model parameters using RMU objectives without involving adversarial perturbations. The RMU loss steers model activations away from harmful knowledge:

$$\mathcal{L}_{RMU} = E_{x_f \sim D_{forget}} \|\mathbf{a}_\ell(x_f) - \mathbf{c}\|_2^2 \quad (6)$$

where $\mathbf{a}_\ell(x_f)$ represents the activation at RMU layer ℓ for forget data, and $\mathbf{c}$ is a random control vector that misdirects the model's internal representations.

To preserve general model capabilities, we employ a retain loss that maintains activations on benign data:

$$\mathcal{L}_{retain} = E_{x_r \sim D_{retain}} \|\mathbf{a}_\ell(x_r) - \mathbf{a}_\ell^{frozen}(x_r)\|_2^2 \quad (7)$$

where $\mathbf{a}_\ell^{frozen}(x_r)$ is the frozen reference model's activation on retain data.

Stage 2: Head-level Perturbation Optimization In the second stage, we optimize the adversarial perturbation parameters while keeping model parameters fixed. The memory preservation loss creates controlled resistance to relearning:

$$\mathcal{L}_{memory} = E_{x_f \sim D_{forget}} MSE(p_\theta(y|x_f), p_{\theta_{ref}}(y|x_f)) \quad (8)$$

Robust Unlearning Mechanism The key insight behind our memory perturbation is that it forces the model to search for robust unlearning directions within a perturbation-radius norm ball, rather than a single vulnerable rejection direction. When adversarial memory perturbations are applied to sensitive heads, the model must find unlearning representations that remain stable across the perturbed space:

$$\mathcal{R}_{robust} = \{\mathbf{h} \in R^d : \|\mathbf{h} - \mathbf{h}_0\| \leq \epsilon\} \quad (9)$$

where $\mathbf{h}_0$ is the original unlearning direction and ϵ is determined by the perturbation magnitude. This approach creates distributed resistance patterns across multiple attention heads, making it significantly more difficult for fine-tuning attacks to recover the original harmful knowledge through simple parameter adjustments.

The alternating optimization between layer-level and head-level parameters ensures that both the global unlearning objective and local robustness constraints are satisfied, resulting in models that maintain strong unlearning effectiveness while being inherently resistant to relearning attempts.

Experiments

In this section, we present our experimental setup and the results of our evaluation, focusing on the effectiveness of our proposed ALMPU method against fine-tuning attacks.

Experimental Setup

Dataset We evaluate our method on the WMDP (Weapons of Mass Destruction Proxy) benchmark(Li et al. 2024). This benchmark consists of multiple-choice questions designed to measure hazardous knowledge in language models. For our unlearning task, we focus on the biosecurity subset (WMDP-Bio), which contains questions about domains such as bioweapons and enhanced potential pandemic pathogens.

For the forget set, we use wmdp-bio-forget-corpus derived from the WMDP-Bio benchmark. For the retain set, we use wiki-text and wmdp-bio-retain-corpus, which helps maintain general and special language modeling capabilities.

Models and Baselines We perform our experiments using Zephyr-7B as the base model. We compare our ALMPU approach against the following baselines: RMU (Representation Misdirection for Unlearning), DPO (Direct Preference Optimization), NPO (Negative Preference Optimization), and we apply our memory perturbation to all layers on a simple unlearn methods GA as an additional baselines GA-MP.

Evaluation Protocol To evaluate the robustness of unlearning methods against adversarial attacks, we perform fine-tuning on different numbers of examples from the forget set after unlearning. For each method and each number of

Algorithm 1: Adaptive Localized Memory Perturbation Unlearning (ALMPU)

Require: Model θ, forget data D_{forget}, retain data D_{retain}
Ensure: Unlearned model θ^* with adversarial perturbations $\{\delta_{l,h}\}$

Phase 1: Adaptive Attention Head Selection
Initialize scaling factors $A_i^l \sim \mathcal{N}(0, \sigma_A)$ for each attention head (l, i)
for each training step in head selection **do**
  Forward pass with scaled attention: $z_t^{(l,i)} \leftarrow (1 + A_i^l) \odot z_t^{(l,i)}$
  Compute selection loss: $\mathcal{L}_{selection} = \mathcal{L}_{CE} + \lambda \sum_{l,i} \|A_i^l\|_1$
  Update scaling factors: $A_i^l \leftarrow A_i^l - \alpha \nabla_{A_i^l} \mathcal{L}_{selection}$
end for
Select top-K heads: $\mathcal{S} = \{(l, i) : \|A_i^l\|_2 intop - K\}$
Phase 2: Alternating Memory Preservation Training
Initialize perturbations: $\delta_{l,h} \sim \mathcal{N}(0, \sigma_{adv}^2 \mathbf{I})$ for $(l, h) \in \mathcal{S}$
Create frozen reference model: $\theta_{ref} \leftarrow \theta$
for each training epoch **do**
  for each batch (x_f, x_r) from (D_{forget}, D_{retain}) **do**
    Stage 1:
    Compute $\mathcal{L}_{layer} = \lambda_{rmu} \|\mathbf{a}_\ell(x_f) - \mathbf{c}\|_2^2 + \lambda_{retain} \|\mathbf{a}_\ell(x_r) - \mathbf{a}_\ell^{frozen}(x_r)\|_2^2$
    Update model parameters: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{layer}$
    Stage 2:
    Apply perturbations $\tilde{\mathbf{h}}_{l,h} = \mathbf{h}_{l,h} + \delta_{l,h}$
    Compute $\mathcal{L}_{head} = \lambda_{memory} MSE(p_\theta, p_{\theta_{ref}}) + \lambda_{reg} \sum \|\delta_{l,h}\|_2^2$
    Update perturbations: $\delta_{l,h} \leftarrow \delta_{l,h} - \eta_\delta \nabla_{\delta_{l,h}} \mathcal{L}_{head}$
  end for
end for
return Unlearned model θ^*

fine-tuning samples, we measure the unlearning effect (UE), defined as:

$$UE = 100\% - accuracy\ on\ D_{forget} \tag{10}$$

A higher UE indicates more successful unlearning. The baseline UE for the original model without any unlearning is 35.4%.

Main Results

Robustness Against Fine-tuning Attacks Table 1 presents the unlearning effect (UE) of different methods before and after fine-tuning on varying numbers of examples from the forget set.

The results show that our ALMPU method significantly outperforms all baselines in terms of robustness against fine-tuning attacks. Because GA is a more aggressive strategy, it achieves the highest initial unlearning effect. However, the robustness against fine-tuning attacks varies across methods. RMU rapidly deteriorates to 51.6% after fine-tuning on just 10 samples. DPO shows similar vulnerability, dropping to 45.5% after 10 samples of fine-tuning. NPO demonstrates better robustness than RMU and DPO but still shows significant degradation after fine-tuning with 10 samples. GA shows rather poor robustness, dropping to 39.2% after 10 samples. Our GA-MP, which incorporates memory perturbation with gradient ascent, shows improved robustness. Most importantly, our ALMPU method demonstrates better robustness, maintaining 70.9% UE after 5 samples and 66.7% after 10 samples, showing minimal degradation compared to the initial unlearning effect. While ALMPU starts with a moderate initial effectiveness (70.1%), it maintains exceptional stability throughout the fine-tuning attack scenarios, outperforming all baseline methods in terms of robustness.

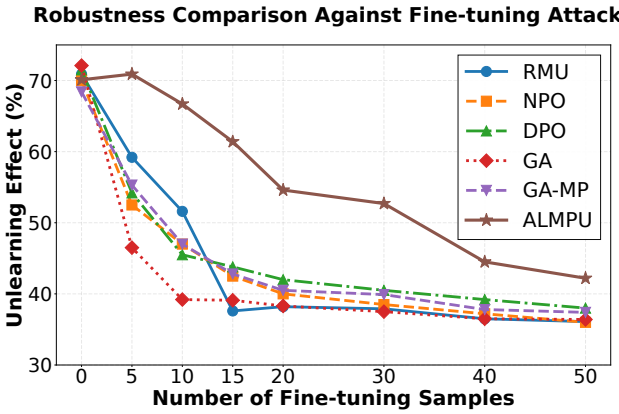

Figure 3: Robustness comparison of different unlearning methods against fine-tuning attacks on WMDP-Bio

Our results show that RMU demonstrate stronger resistance to fine-tuning attacks compared to the original adversarial safety paper (Łucki et al. 2024), which can be attributed to differences in experimental methodology. The original evaluation used a partially disclosed multiple-choice dataset (wmdp-bio-dataset_mc) for robustness testing, while our evaluation employs the publicly available WMDP-Bio corpus (wmdp-bio-corpora) for fine-tuning attacks. This choice ensures both accessibility and data format consistency, as the original approach involved training on LLM-summarized multiple-choice datasets that differ from RMU's original corpora training. Our corpus-based evaluation provides a fair comparison framework where all methods use the same data format as they were trained.

Attention Head Sensitivity Analysis To better understand how our adaptive attention head selection mechanism identifies the most sensitive components for harmful knowledge processing, we visualize the learned scaling factors across all attention heads in the Zephyr-7B model. Figure 4 presents a heatmap of attention head sensitivity scores, where each cell represents the L2 norm of the learned scaling factors $\|A_i^l\|_2$ for head i at layer l after normalization.

We observe that sensitivity is not uniformly distributed across layers, with certain layers (particularly layers 5-9) showing concentrated areas of high sensitivity which is supported by the results of other work(Huu-Tien et al. 2024).

Method	0 sample	5 samples	10 samples	15 samples	20 samples	30 samples	40 samples	50 samples
RMU	70.9	59.2	51.6	37.6	38.2	37.9	36.5	36.1
NPO	70.0	52.5	47.0	42.5	40.0	38.5	37.2	36.0
DPO	71.5	54.2	45.5	43.8	42.0	40.5	39.2	38.0
GA	**72.1**	46.5	39.2	39.1	38.3	37.5	36.5	36.4
GA-MP	68.4	55.3	47.0	42.8	40.5	39.9	37.8	37.4
ALMPU	70.1	**70.9**	**66.7**	**61.4**	**54.6**	**52.7**	**44.5**	**42.2**

Table 1: Robustness evaluation of unlearning methods against fine-tuning attacks on WMDP-Bio

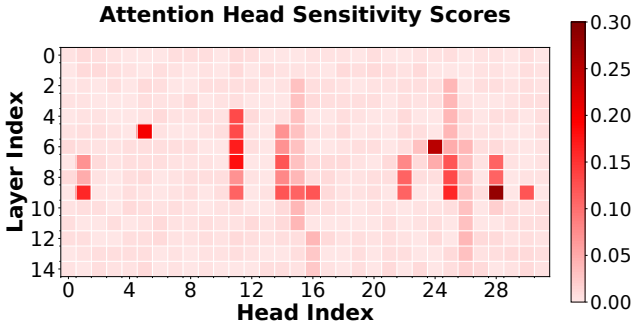

Figure 4: Attention head sensitivity score heatmap. Darker red colors indicate higher sensitivity scores.

This suggests that harmful knowledge processing is localized to specific computational stages within the transformer architecture. Notably, we identify several attention heads with particularly high sensitivity scores. These heads represent the most critical components for harmful knowledge processing and are prioritized in our top-K selection for targeted intervention. The sparse distribution of high-sensitivity heads validates our approach of selective intervention rather than uniform modifications across all attention components.

This analysis provides empirical evidence for the effectiveness of our adaptive selection mechanism in identifying genuinely important components for unlearning. The clear localization patterns also suggest that our method successfully captures the mechanistic structure of how harmful knowledge is represented and processed within large language models.

Logits Lens Analysis To examine how unlearning methods affect internal model representations, we conduct a logits lens analysis across all 32 transformer blocks by evaluating WMDP-Bio accuracy at each intermediate layer, where lower values indicate better unlearning effectiveness. The vanilla baseline exhibits consistently high accuracy with a notable spike in blocks 17-22. NPO and DPO demonstrate partial suppression, maintaining accuracy around 0.3-0.4 in early blocks but showing elevated accuracy in the knowledge-rich region. RMU demonstrates more effective suppression, maintaining accuracy close to random chance throughout most blocks with only modest elevation in critical regions. Our ALMPU method achieves similar performance to RMU, maintaining accuracy consistently near random chance across all transformer blocks. The close alignment between ALMPU and RMU validates that both representation-based methods effectively suppress harmful representations.

Wikitext results To further evaluate the robustness of different unlearning methods, we conduct fine-tuning experiments using the Wiki-text dataset, which contains general knowledge unrelated to the harmful content being unlearned. This experiment simulates realistic attack scenarios where adversaries use benign, publicly available data to compromise unlearning effectiveness. Table 2 presents the unlearning effect (UE) after fine-tuning with different amounts of Wiki-text data.

Method	10	50	100	200	500
RMU	**70.1**	69.2	**70.0**	**69.2**	**68.5**
NPO	58.2	49.2	42.6	40.3	39.8
ALMPU	69.5	**69.3**	69.2	68.6	68.1

Table 2: Unlearning effect on Wiki-text dataset across different size of fine-tuning samples

The results reveal significant differences in robustness across unlearning methods when faced with unrelated data fine-tuning. present in Wiki-text training.

NPO shows vulnerability to this type of attack, with unlearning effectiveness degrading dramatically as fine-tuning samples increase. The drop from 58.2% to 39.8% represents a 32% relative decrease in effectiveness, indicating that NPO's probability-based optimization is highly susceptible to being reversed by exposure to natural language data. This vulnerability stems from NPO's reliance on preference learning objectives that can be easily overwhelmed by standard language modeling losses during fine-tuning.

RMU demonstrates strong robustness, maintaining consistently high unlearning effectiveness (68.5-70.1%) across all fine-tuning intensities. This stability suggests that RMU's representation-based approach creates modifications that are relatively resistant to being overwritten by general language modeling objectives. Our ALMPU method demonstrates similar robustness compared to RMU, maintaining unlearning effectiveness between 68.1% and 69.9% across all conditions.

Ablation Studies

Impact of Attention Head Selection We investigate the effectiveness of our adaptive attention head selection mechanism by comparing different head selection strategies. Table

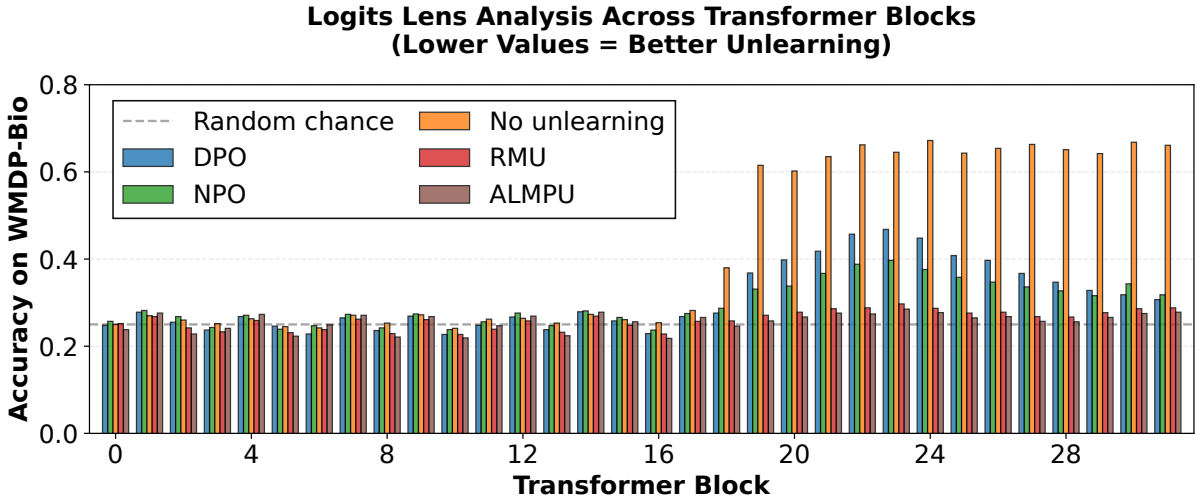

Figure 5: Logits lens analysis showing accuracy on WMDP-Bio across transformer blocks. Lower values indicate better unlearning effectiveness at the representational level.

3 presents the results with random head selection at different percentages, full head selection, and our adaptive scaling factor approach. All attention heads are from the same layer as RMU chooses.

Strategy	0 sample	25 samples	50 samples
Ours	**70.1**	53.4	42.2
Random 25%	69.2	45.1	38.2
Random 50%	67.8	48.9	40.1
Random 75%	64.3	52.2	40.3
Full	63.9	**54.8**	**43.1**

Table 3: Comparison of different attention head selection strategies' influences on UE.

The results demonstrate the importance of strategic attention head selection in our adaptive mechanism. Random 25% head selection shows limited performance. As we increase the percentage of randomly selected heads, we observe mixed results: Random 50% achieves moderate performance with 67.8% initial effectiveness, while Random 75% shows improved robustness at 25 samples but reduced initial effectiveness. Full head selection strategy demonstrates strong robustness, achieving the best performance at 25 and 50 samples, but suffers from reduced initial unlearning effectiveness. This suggests that while broader coverage provides robustness benefits, it may interfere with the initial unlearning process. Our adaptive scaling factor approach achieves the optimal balance, maintaining the highest initial unlearning effect while demonstrating competitive robustness performance.

The location-based ablation study further validates our selective perturbation strategy. Among fixed layer approaches, Late Layers achieve the highest initial unlearning effect but show poor robustness after 25 samples. Middle Layers demonstrate more balanced performance and better robustness. Early Layers show the most consistent but overall

Location	0 sample	25 samples	50 samples
Early layers	65.1	47.2	37.8
Mid layers	68.2	49.8	39.2
Late layers	69.4	41.5	36.4
All layers	64.5	50.8	41.1
Ours	**70.1**	**53.4**	**42.2**

Table 4: Comparison of memory preservation perturbation locations

weaker performance across all metrics.The All Layers approach, while providing reasonable robustness, suffers from reduced initial effectiveness, confirming that uniform perturbation introduces unnecessary interference. Our adaptive approach consistently outperforms all fixed strategies, achieving the highest initial unlearning effect and maintaining superior robustness. This validates our hypothesis that harmful knowledge processing is concentrated in specific computational components rather than uniformly distributed, and that precise targeting of these sensitive locations through our adaptive mechanism yields optimal results.

Conclusion

In this paper, we introduced ALMPU (Adaptive Localized Memory Perturbation Unlearning), a novel approach to machine unlearning that addresses the limitations of existing methods through adaptive knowledge guidance and memory perturbation mechanisms. Our method introduces two key innovations: an adaptive attention head selection mechanism that identifies the most sensitive components for harmful knowledge processing through learnable scaling factors, and a memory perturbation defense that makes inherently resistant to relearning attempts. Experimental results on the WMDP benchmark demonstrate that ALMPU outperforms existing methods in terms of fine-tuning attack robustness.

Acknowledgements

This study was supported by National Natural Science Foundation of China No.62176252.

References

Barez, F.; Fu, T.; Prabhu, A.; Casper, S.; Sanyal, A.; Bibi, A.; O'Gara, A.; Kirk, R.; Bucknall, B.; Fist, T.; et al. 2025. Open problems in machine unlearning for ai safety. *arXiv preprint arXiv:2501.04952*.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Ding, C.; Wu, J.; Yuan, Y.; Lu, J.; Zhang, K.; Su, A.; Wang, X.; and He, X. 2025. Unified Parameter-Efficient Unlearning for LLMs. arXiv:2412.00383.

Feng, Z.; Xu, Y. E.; Robey, A.; Kirk, R.; Davies, X.; Gal, Y.; Schwarzschild, A.; and Kolter, J. Z. 2025. Existing Large Language Model Unlearning Evaluations Are Inconclusive. arXiv:2506.00688.

Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2023. A framework for few-shot language model evaluation.

Huu-Tien, D.; Pham, T.-T.; Thanh-Tung, H.; and Inoue, N. 2024. On effects of steering latent representation for large language model unlearning. *arXiv preprint arXiv:2408.06223*.

Jia, J.; Zhang, Y.; Zhang, Y.; Liu, J.; Runwal, B.; Diffenderfer, J.; Kailkhura, B.; and Liu, S. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Jin, Z.; Cao, P.; Wang, C.; He, Z.; Yuan, H.; Li, J.; Chen, Y.; Liu, K.; and Zhao, J. 2024. RWKU: Benchmarking Real-World Knowledge Unlearning for Large Language Models. *arXiv preprint arXiv:2406.10890*.

Lermen, S.; Rogers-Smith, C.; and Ladish, J. 2023. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Phan, L.; et al. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Liu, Z.; Dou, G.; Tan, Z.; Tian, Y.; and Jiang, M. 2024. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*.

Liu, Z.; Ye, H.; Chen, C.; Zheng, Y.; and Lam, K.-Y. 2025. Threats, attacks, and defenses in machine unlearning: A survey. *IEEE Open Journal of the Computer Society*.

Łucki, J.; Wei, B.; Huang, Y.; Henderson, P.; Tramèr, F.; and Rando, J. 2024. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Shi, W.; Lee, J.; Huang, Y.; Malladi, S.; Zhao, J.; Holtzman, A.; Liu, D.; Zettlemoyer, L.; Smith, N. A.; and Zhang, C. 2024. MUSE: Machine Unlearning Six-Way Evaluation for Language Models. arXiv:2407.06460.

Shumailov, I.; Hayes, J.; Triantafillou, E.; Ortiz-Jimenez, G.; Papernot, N.; Jagielski, M.; Yona, I.; Howard, H.; and Bagdasaryan, E. 2024a. UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI. arXiv:2407.00106.

Shumailov, I.; Hayes, J.; Triantafillou, E.; Ortiz-Jimenez, G.; Papernot, N.; Jagielski, M.; Yona, I.; Howard, H.; and Bagdasaryan, E. 2024b. UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI. *arXiv preprint arXiv:2407.00106*.

Suriyakumar, V. M.; Sekhari, A.; and Wilson, A. 2025. UCD: Unlearning in LLMs via Contrastive Decoding. arXiv:2506.12097.

Vasilev, S.; Herold, C.; Liao, B.; Hashemi, S. H.; Khadivi, S.; and Monz, C. 2025. Unilogit: Robust Machine Unlearning for LLMs Using Uniform-Target Self-Distillation. arXiv:2505.06027.

Yao, J.; Chien, E.; Du, M.; Niu, X.; Wang, T.; Cheng, Z.; and Yue, X. 2024. Machine Unlearning of Pre-trained Large Language Models. arXiv:2402.15159.

Yao, Y.; Xu, X.; and Liu, Y. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Yin, F.; Ye, X.; and Durrett, G. 2024. Lofit: Localized fine-tuning on llm representations. *Advances in Neural Information Processing Systems*, 37: 9474–9506.

Yuan, H.; Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2025. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25769–25777.

Yuan, X.; Pang, T.; Du, C.; Chen, K.; Zhang, W.; and Lin, M. 2024. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Zhong, X.; Luo, H.; and Liu, C. 2025. DualOptim: Enhancing Efficacy and Stability in Machine Unlearning with Dual Optimizers. arXiv:2504.15827.