Representation-Guided Parameter-Efficient LLM Unlearning

Zeguan Xiao[1*], **Lang Mo**[2*], **Yun Chen**[1], **Lei Yang**[3], **Jiehui Zhao**[3]
Lili Yang[2†], **Guanhua Chen**[2†]
[1]Shanghai University of Finance and Economics
[2]Southern University of Science and Technology, [3]Deepexi Technology Co. Ltd.

Abstract

Large Language Models (LLMs) often memorize sensitive or harmful information, necessitating effective machine unlearning techniques. While existing parameter-efficient unlearning methods have shown promise, they still struggle with the forget-retain trade-off. This can be attributed to their reliance on parameter importance metrics to identify parameters that are important exclusively for forget set, which is fundamentally limited by the superposition phenomenon. Due to the polysemantic nature of LLMs parameters, such an importance metric may struggle to disentangle parameters associated with forget and retain sets. In this work, we propose **Re**presentation-**G**uided **L**ow-rank **U**nlearning (ReGLU), a novel approach that leverages the geometric properties of representation spaces to achieve robust and precise unlearning. First, we develop a representation-guided initialization for LoRA that identifies the optimal subspace for selective forgetting. Second, we introduce a regularization loss that constrains the outputs of the LoRA update to lie in the orthogonal complement of the retain set's representation subspace, thereby minimizing interference with the model's performance on the retain set. We evaluate ReGLU on the TOFU and WMDP benchmarks across multiple models. Our results demonstrate that ReGLU consistently outperforms state-of-the-art baselines, achieving superior unlearning quality while maintaining higher model utility.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across various natural language processing tasks, demonstrating emergent abilities in reasoning, coding, and creative writing (Wei et al., 2022). However, these models are often trained on massive, uncurated datasets that may contain sensitive personal information, copyrighted content, or harmful biases (Brown et al., 2022; Bender et al., 2021). The tendency of LLMs to memorize and regenerate such data poses significant risks to privacy and intellectual property (Carlini et al., 2022; Nasr et al., 2023). Consequently, there is an urgent need for Machine Unlearning (MU) (Cao and Yang, 2015), which aims to remove specific knowledge from a pre-trained model without the prohibitive cost of retraining from scratch.

Most existing LLM unlearning methods rely on full fine-tuning, i.e., updating all model parameters. However, modifying billions of parameters is computationally expensive and, more critically, substantially increases the risk of catastrophic forgetting. To address this issue, recent studies have leveraged LoRA (Hu et al., 2022) for LLM unlearning and demonstrated that unlearning performance can be comparable to, or even better than, full fine-tuning while substantially reducing computational cost (Cha et al., 2025; Kim et al., 2025).

Despite this progress, these methods still struggle with the forget-retain trade-off: reducing performance on the forget set often comes at the cost of degraded performance on the retain set (Fisher, 1922; Cha et al., 2025; Kim et al., 2025; Xiao et al., 2026). We hypothesize that this limitation stems from their reliance on estimating parameter importance (e.g., via Fisher information (Fisher, 1922)) while neglecting the phenomenon of superposition (Elhage et al., 2022) in LLMs. The superposition phenomenon implies that a single parameter is often involved in the representation of multiple concepts, leading to polysemanticity of parameters. Consequently, relying on such importance measures to isolate forget-related parameters becomes problematic, as these parameters are often simultaneously crucial for maintaining performance on the retain set or other unseen general knowledge.

Our approach is built on a key insight: while parameter importance measures may be unreliable due to superposition, the representation subspaces

[*] Equal Contribution.
[†]Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 14602–14616

can be more effectively disentangled (Zou et al., 2023). By constraining unlearning updates within a subspace that aligns with forget-set representations while minimizing interference with retain-set representations, we can more effectively isolate forget-related knowledge and preserve model utility.

In this work, we propose **Re**presentation-**G**uided **L**ow-rank **U**nlearning (ReGLU), a novel approach that leverages the geometric properties of representation spaces to achieve robust and precise unlearning. Our approach consists of two components. First, we develop RILA (**R**epresentation-guided **I**nitialization of **L**ow-rank **A**daption), a LoRA initialization strategy for LLM unlearning. RILA identifies a balanced subspace that maximizes the variance of forget-set representations while minimizing that of the retain set. Second, we introduce ROL (**R**epresentation **O**rthogonal **L**oss), a regularization loss term for unlearning. By identifying the principal subspace of the retain set's representations, this loss enforces an orthogonality constraint on the LoRA up-projection matrices. This ensures that the LoRA update lies in the orthogonal complement of a subspace of the representations of retain set, thereby minimizing the interference with the original model's behavior.

We evaluate ReGLU on two widely used LLM unlearning benchmarks: TOFU (Maini et al., 2024) and WMDP (Li et al., 2024). Our results demonstrate that ReGLU consistently outperforms baselines.

Our contributions are summarized as follows[1]:

- **Methodology:** We propose ReGLU, a novel framework that shifts the LoRA-based LLM unlearning paradigm from parameter importance to representation geometry.

- **Experiments:** We conduct extensive evaluations on the TOFU and WMDP benchmarks across multiple models, including Llama-2-7B (Touvron et al., 2023), Phi-1.5 (Li et al., 2023), and Zephyr-7B-beta (Tunstall et al., 2024). The results demonstrate that ReGLU consistently establishes new state-of-the-art performance.

- **Analysis:** We provide a theoretical guarantee for our initialization strategy and offer in-depth geometric diagnostics. Our analysis confirms that ReGLU successfully disentangles forget and retain representations, validating that subspace-level control is more robust than parameter-level estimation for unlearning.

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

2 Background

2.1 Problem Definition

For a model $\mathcal{M}$ that is 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 only trained on $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$, referred to as the *retain set*.

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

2.2 Low-Rank Adaptation (LoRA)

LoRA (Hu et al., 2022) is a parameter-efficient fine-tuning technique that allows effective adaptation with significantly fewer trainable parameters. Specifically, for a linear layer with weight matrix $W_0 \in \mathbb{R}^{d_{out} \times d_{in}}$, LoRA parameterizes the update as a low-rank update ΔW:

$$W = W_0 + \Delta W = W_0 + BA,$$

where $B \in \mathbb{R}^{d_{out} \times r}$ and $A \in \mathbb{R}^{r \times d_{in}}$ are the low-rank matrices with $r \ll \min(d_{out}, d_{in})$. During finetuning, only the matrices B and A are updated, while the original weights W_0 remain fixed.

2.3 LLM Unlearning Methods

Given a pre-trained LLM with parameters $\boldsymbol{\theta}$, we denote the probability distribution it defined as $p(x; \boldsymbol{\theta})$, where x represents the input text.

The most prevalent approach to unlearning is to suppress the model's likelihood on the forget set $\mathcal{D}_f$—that is, drive $p(x; \boldsymbol{\theta})$ downward for $x \in \mathcal{D}_f$. This is commonly 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.

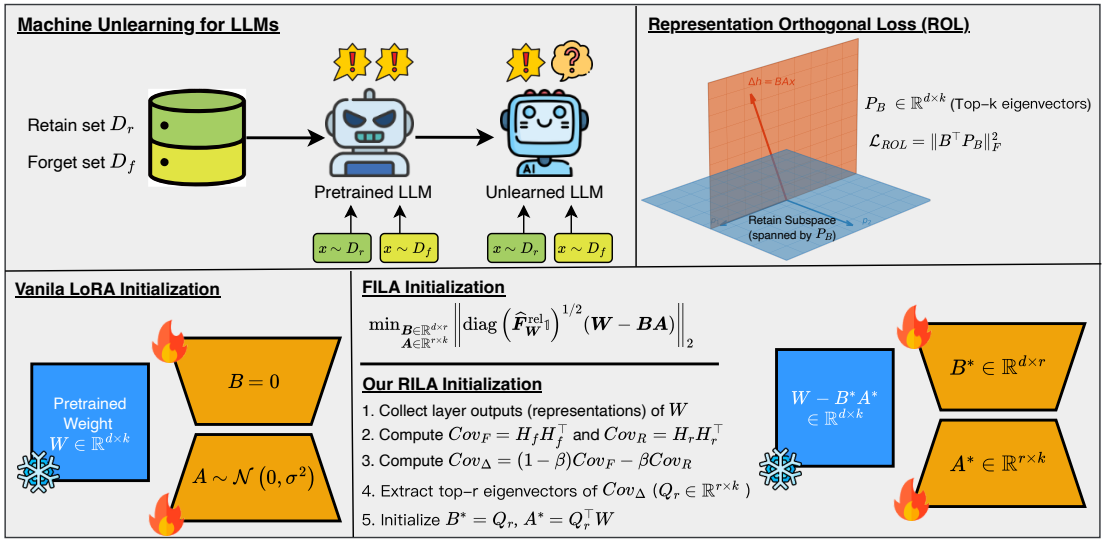

Figure 1: LLM unlearning aims to remove specific information from a pre-trained model. FILA estimates parameter importances $\widehat{F}_W^{\mathrm{rel}}$ and solves a weighted low-rank approximation problem to initialize LoRA matrices. Our ReGLU framework collects layer representations and leverages representation geometry to guide selective forgetting while preserving retain set knowledge.

Given the inherent issues of GA (Zhang et al., 2024; Cha et al., 2025), some methods have been proposed as alternative unlearning objectives. For instance, 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, allowing for more controlled divergence during unlearning, preventing catastrophic collapse.

Inverted Hinge Loss (IHL) (Cha et al., 2025) is another approach designed to simultaneously address the issues of unbounded loss, gradient spread, and the degradation of generative performance:

$$\mathcal{L}_{\mathrm{IHL}}(\boldsymbol{x}) = 1 + p\left(x_t|x_{<t};\boldsymbol{\theta}\right) - \max_{v\neq x_t} p\left(v|x_{<t};\boldsymbol{\theta}\right).$$

FILA (Cha et al., 2025) and VILA (Kim et al., 2025) employ Fisher Information (FI) to identify parameters associated with a dataset. The FI of a dataset $\mathcal{D}$ with respect to model parameters $\boldsymbol{\theta}$ is defined as:

$$\mathcal{F}_{\boldsymbol{\theta}}(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(x;\boldsymbol{\theta})\right)^2\right].$$

FILA computes the ratio of FI values from the forget set and retain set to determine the *forget importance map* $\mathcal{S} = \mathcal{F}_{\boldsymbol{\theta}}(\mathcal{D}_f)/\mathcal{F}_{\boldsymbol{\theta}}(\mathcal{D}_r)$. By leveraging this map, FILA initializes the LoRA matrices A and B such that BA captures the components of the original weight matrix W_0 most relevant to the forget set. VILA extends FILA by addressing the inaccuracy in FI as a variance measure when the score function has a non-zero expectation.

3 Methodology

In this section, we present ReGLU, a representation-guided framework for LoRA-based LLM unlearning. Our framework consists of two complementary components. First, in Section 3.1, we introduce RILA, a LoRA initialization strategy that leverages the geometric structure of representation spaces to align the initial LoRA update with a subspace maximally discriminative between the forget and retain sets. Second, in Section 3.2, we propose ROL, a regularization loss that enforces orthogonality between the LoRA update and the principal subspace of retain-set representations, preventing interference with preserved knowledge during training. We conclude with the overall optimization objective that combines these components.

3.1 RILA: A Representation-Guided LoRA Initialization for LLM Unlearning

3.1.1 Motivation

Consider a linear layer $h = W_0 x$, where $x \in \mathbb{R}^{d_{in}}$ is the input representation, $W_0 \in \mathbb{R}^{d_{out}\times d_{in}}$ is the pre-trained weight matrix, and $h \in \mathbb{R}^{d_{out}}$ is the

output representation. Fine-tuning for unlearning aims to learn a parameter update ΔW such that the updated weight $W = W_0 + \Delta W$ reduces the likelihood of the forget set $\mathcal{D}_f$ while maintaining performance on the retain set $\mathcal{D}_r$. Ideally, the update ΔW should have minimal impact on the outputs for inputs from the retain set. This can be expressed as $(W_0 + \Delta W)x \approx W_0 x$, which implies $\Delta W x \approx 0$ for all $x \in \mathcal{D}_r$. For the forget set $\mathcal{D}_f$, the update ΔW should produce a substantial change in the output to effectively suppress the model's ability to recall target knowledge.

To translate these requirements into a formal objective, we define the unlearning problem as finding an update ΔW that maximizes the "differential energy" between the forget and retain sets. Specifically, we aim to maximize the expected change in output norm for the forget set while simultaneously minimizing it for the retain set:

$$\max_{\Delta W} \quad (1-\beta)\,\mathbb{E}_{x\sim\mathcal{D}_f}\left[\|\Delta W x\|_2^2\right] - \beta\,\mathbb{E}_{x\sim\mathcal{D}_r}\left[\|\Delta W x\|_2^2\right], \tag{1}$$

where $\beta \in [0, 1]$ is a hyperparameter that balances the trade-off between forget and retain.

LoRA-based LLM unlearning methods, FILA (Cha et al., 2025) and VILA (Kim et al., 2025), have achieved promising results by leveraging informed initialization strategies. Meanwhile, recent research on LoRA (Meng et al., 2024; Wang et al., 2025) highlights that the initialization strategy plays a crucial role in determining the convergence behavior and final performance of the model. Inspired by these successes, we directly model the weight update ΔW as LoRA and develop our LoRA initialization strategy to solve the unlearning problem by initializing the LoRA matrices to explicitly align with the differential energy objective defined in Eq. 1.

3.1.2 Representation-Guided Initialization

Following the motivation in the previous section, we aim to initialize the LoRA matrices A and B such that the update $\Delta W = BA$ is initially aligned with a subspace that maximizes the impact on the forget set $\mathcal{D}_f$ while minimizing the interference with the retain set $\mathcal{D}_r$.

Our key insight is to leverage the decomposability of representations (Zou et al., 2023) to guide the initialization of LoRA. This approach circumvents the issues of parameter polysemanticity caused by the superposition phenomenon (Elhage et al., 2022), which can be a potential limitation when directly employing parameter importance to initialize LoRA matrices.

We formalize our initialization strategy and its theoretical justification in the following theorem:

Theorem 1. *Consider the distributions $\mathcal{P}_F$ and $\mathcal{P}_R$ of the output representation $h = W_0 x$ for inputs x sampled from $\mathcal{D}_f$ and $\mathcal{D}_r$, respectively. Let Cov_F and Cov_R be their corresponding covariance matrices:*

$$\mathrm{Cov}_F = \mathbb{E}_{h\sim\mathcal{P}_F}[hh^\top], \quad \mathrm{Cov}_R = \mathbb{E}_{h\sim\mathcal{P}_R}[hh^\top].$$

Define Cov_Δ as:

$$\mathrm{Cov}_\Delta = (1-\beta)\,\mathrm{Cov}_F - \beta\,\mathrm{Cov}_R,$$

where $\beta \in [0, 1]$ is a hyperparameter. When A, B are initialized as $B_{init} = Q_r$ and $A_{init} = Q_r^\top W_0$, where $Q_r = [q_1, q_2, \ldots, q_r]$ is the matrix of the top-r eigenvectors of Cov_Δ, Eq. 1 is maximized at initialization.

The proof of Theorem 1 is provided in Appendix A. Intuitively, this theorem reveals that by constructing a balanced covariance matrix Cov_Δ that contrasts the forget and retain sets, we can extract eigenvectors that capture the most discriminative directions. These eigenvectors define a subspace where the forget-set representations have high variance while the retain-set representations have low variance, naturally aligning with our goal of selective forgetting.

In practice, since the true distributions $\mathcal{P}_F$ and $\mathcal{P}_R$ are unknown, we estimate Cov_F and Cov_R empirically: feed a small number of samples from $\mathcal{D}_f$ and $\mathcal{D}_r$ through the pre-trained model, collect layer outputs $H_f \in \mathbb{R}^{N_f \times d}$ and $H_r \in \mathbb{R}^{N_r \times d}$, and compute the sample covariances. Using these, build Cov_Δ and extract Q_r to perform the initialization above. Appendix B presents a concentration bound for empirical covariance estimation, showing that under bounded activations the empirical estimators converge in spectral norm at the rate $O(M^2\sqrt{\log(d/\delta)/N})$, where M is an upper bound on the activation norm, d is the representation dimension, N is the number of samples, and δ controls the confidence level of the bound. As a result, the empirical balanced covariance remains close to Cov_Δ, and its top-r eigenspace is stable whenever the eigengap of Cov_Δ is sufficiently large.

3.2 Representation Orthogonal Loss: A Subspace-Controlled Regularization

To further ensure that the unlearning process maintains performance on the retain set $\mathcal{D}_r$, we introduce Representation Orthogonal Loss (ROL), a subspace-controlled regularization term in training loss. This loss is designed to constrain the output of the LoRA to be orthogonal to the retain set's representations, thereby minimizing interference with the knowledge that should be preserved.

We first identify the principal subspace of the representations for the retain set $\mathcal{D}_r$. Let $H_r = \{h_i\}_{i=1}^N$ be the set of output representations $h = W_0 x$ collected from a small subset of the retain set. We perform eigenvalue decomposition on the covariance matrix of H_r to obtain an orthonormal basis $P_B \in \mathbb{R}^{d_{out} \times k}$, where k is a hyperparameter controls the strength of the regularization. This matrix P_B captures the most critical directions in the representation space that represent the knowledge to be retained.

During training, P_B remains fixed. We define the ROL as:

$$\mathcal{L}_{ROL} = \|\mathbf{B}^\top P_B\|_F^2, \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{d_{out} \times r}$ is the LoRA up-projection matrix and $\|\cdot\|_F$ denotes the Frobenius norm. This loss function regularizes the orthogonality between every pair of columns of $\mathbf{B}$ and P_B. Geometrically, this loss encourages that the LoRA update $\Delta h = B(Ax)$ lies in the orthogonal complement of a subspace of the representations of retain set, thereby minimizing the interference with the original model's behavior on $\mathcal{D}_r$.

3.3 Overall Algorithm

The final optimization objective for ReGLU is a weighted combination of the forget loss, the retain loss, and the orthogonal regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forget}}(\mathcal{D}_f) + \gamma\, \mathcal{L}_{\text{retain}}(\mathcal{D}_r) + \lambda \mathcal{L}_{ROL},$$

where $\mathcal{L}_{\text{forget}}$ can be any of the objectives discussed in Section 2.3, and γ, λ are hyperparameters balancing the different objectives.

The overall unlearning process of ReGLU is summarized in Algorithm 1 in Appendix C. First, we collect a small number of samples from both the forget set $\mathcal{D}_f$ and the retain set $\mathcal{D}_r$ to estimate the covariance matrices Cov_F and Cov_R. We then compute the balanced covariance matrix Cov_Δ and extract its top-r eigenvectors to initialize the LoRA matrices A and B as described in Section 4.2. Simultaneously, we use Cov_R to construct the orthogonal basis P_B. Finally, we optimize the LoRA parameters by minimizing the total loss $\mathcal{L}_{\text{total}}$, which balances unlearning effectiveness, retain set performance, and subspace-controlled regularization.

4 Experiments

4.1 Experimental Setup

Benchmarks. We conduct experiments on two widely used LLM unlearning benchmarks: TOFU (Maini et al., 2024) and WMDP (Li et al., 2024). TOFU offers 200 diverse synthetic author profiles, each consisting of 20 question-answer pairs. Subsets of these profiles (1%, 5%, or 10%) serve as the forget set for unlearning. 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.

Baselines. We compare our method with other LoRA-based unlearning methods, specifically FILA (Cha et al., 2025) and VILA (Kim et al., 2025), with two unlearning loss functions: GD (Liu et al., 2022) and IHL (Cha et al., 2025). We focus our evaluation on LoRA-based approaches to ensure a fair comparison within the parameter-efficient framework. This choice is motivated by findings in Cha et al. (2025), which demonstrate that LoRA-based unlearning can achieve performance on par with, or even superior to, full fine-tuning.

Models. We follow the default configurations for each benchmark. For TOFU, we evaluate unlearning on Llama-2-7B (Touvron et al., 2023) and Phi-1.5B (Li et al., 2023). For WMDP, we use Zephyr-7B-beta (Tunstall et al., 2024).

Evaluation Metrics and Setting. We follow the standard evaluation protocols for each benchmark. For TOFU, we employ the Forget Quality (FQ) (Maini et al., 2024) to measure the extent of data removal, which is the statistical similarity between the unlearned model and an oracle retrained model. Note that FQ values reported in the results are log-scaled to facilitate visualization and comparison across different magnitudes. The model utility is the harmonic mean of nine metrics (Maini

Model	Method	Forget 1% ↑	Forget 5% ↑	Forget 10% ↑	AVG. ↑
Phi-1.5B	Original Model	-2.5	-11.5	-16.9	-10.3
	GD	-2.5	-11.2	-17.3	-10.3
	GD + FILA	-1.8	-8.7	-11.9	-7.5
	GD + VILA	-2.5	-10.9	-14.4	-9.3
	GD + ReGLU	-0.8	-9.9	-11.2	-7.3
	IHL	-1.3	-11.5	-12.4	-8.4
	IHL + FILA	-2.5	-9.3	-10.3	-7.4
	IHL + VILA	-2.9	-10.2	-10.2	-7.8
	IHL + ReGLU	-0.1	-5.4	-7.7	-4.4
Llama-2-7B	Original	-2.9	-14.0	-16.9	-11.3
	GD	-3.3	-13.2	-13.8	-10.1
	GD + FILA	-1.3	-11.2	-16.6	-9.7
	GD + VILA	-2.5	-13.6	-14.7	-10.3
	GD + ReGLU	-0.6	-12.9	-13.3	-8.9
	IHL	-2.9	-14.3	-16.6	-11.3
	IHL + FILA	-2.2	-11.2	-7.5	-6.9
	IHL + VILA	-2.5	-8.2	-11.1	-7.2
	IHL + ReGLU	-0.7	-5.4	-1.1	-2.4

Table 1: Main results on the TOFU benchmark. We report the log-scaled Forget Quality (FQ) for different forget set sizes (1%, 5%, and 10%) on Phi-1.5B and Llama-2-7B. Higher FQ indicates better unlearning performance, signifying that the unlearned model's behavior is statistically closer to a model retrained from scratch. All methods maintain at least 95% of the original model's utility.

et al., 2024). For WMDP, we report the accuracy of WMDP-Bio and WMDP-Cyber to assess the efficacy of unlearning and evaluate model utility using MMLU accuracy (Hendrycks et al., 2021). To ensure a fair comparison, we search the hyperparameters of all methods and only consider checkpoints that maintain at least 95% of the original model's utility and select the one achieving the best unlearning performance. Since all methods demonstrate comparable utility under this criterion, we focus on reporting the forgetting performance in the subsequent results.

4.2 Main Results

We present the main unlearning results on the TOFU and WMDP benchmarks in Table 1 and Table 2, respectively.

Results on TOFU. As shown in Table 1, ReGLU consistently outperforms the baseline methods across different forget set sizes (1%, 5%, and 10%) and model architectures (Phi-1.5B and Llama-2-7B). When using GD as the loss function, ReGLU achieves an average FQ of -7.3 on Phi-1.5B and -8.9 on Llama-2-7B, surpassing both FILA and VILA. The performance gain is even more pronounced when combined with the IHL. Specifically, IHL + ReGLU achieves the best overall performance, with an average FQ of -4.4 on Phi-1.5B and -2.4 on Llama-2-7B. This demonstrates that our ReGLU effectively complements different unlearning functions.

Results on WMDP. The results on the WMDP benchmark, summarized in Table 2, further validate the effectiveness of ReGLU. GD + ReGLU achieves the lowest average accuracy of 35.1% across the Bio and Cyber domains, which is a significant reduction from the original model's 54.6% and outperforms GD + VILA (38.1%). Similarly, IHL + ReGLU (41.4%) shows a substantial improvement over IHL + FILA (48.8%) and IHL + VILA (51.0%). These results indicate that ReGLU can more aggressively suppress the forget-set knowledge without compromising the model's general capabilities.

Method	Bio ↓	Cyber ↓	AVG. ↓
Original Model	64.7	44.5	54.6
GD	58.0	37.3	47.7
GD + FILA	59.1	41.3	50.2
GD + VILA	49.3	26.9	38.1
GD + ReGLU	41.8	28.4	**35.1**
IHL	54.9	39.5	47.2
IHL + FILA	60.6	37.0	48.8
IHL + VILA	63.6	38.4	51.0
IHL + ReGLU	49.7	33.1	**41.4**

Table 2: Main results on the WMDP benchmark. We report the accuracy (%) on WMDP-Bio and WMDP-Cyber. Lower accuracy indicates better unlearning performance. All reported checkpoints maintain at least 95% of the original model's MMLU utility.

Method	Bio ↓	Cyber ↓	AVG. ↓
GD	58.0	37.3	47.7
GD + RILA	54.0	25.8	39.9
GD + ROL	44.6	36.1	40.4
GD + ReGLU	41.8	28.4	**35.1**
IHL	54.9	39.5	47.2
IHL + RILA	51.2	36.7	44.0
IHL + ROL	54.0	36.1	45.0
IHL + ReGLU	49.7	33.1	**41.4**

Table 3: Ablation results on the WMDP benchmark. We compare the full ReGLU framework with two variants: RILA (representation-guided initialization only) and ROL (subspace regularization only). Lower values indicate better unlearning performance.

4.3 Ablation Studies

To investigate the contribution of each component in ReGLU, we conduct ablation studies on the WMDP benchmark using both GD and IHL. We compare the full ReGLU framework with two variants: (1) RILA, which only employs the representation-guided initialization described in Section 3.1 without the subspace regularization loss ROL; and (2) ROL, which only applies the subspace regularization loss described in Section 3.2 while using standard LoRA initialization. As shown in Table 3, both the initialization and the regularization loss contribute to the overall performance. Specifically, RILA achieves better unlearning performance than the base GD and IHL, demonstrating that aligning the LoRA update with the forget-set-dominant subspace provides a strong starting point for unlearning. Similarly, ROL also shows improvements over the base methods, indicating that the subspace regularization loss effectively guides the optimization process. Most importantly, the full ReGLU framework, which combines both components, achieves the best results. This suggests a strong synergy between the geometric initialization and the structural regularization, where the initialization provides a well-aligned starting subspace and the regularization ensures that the subsequent updates remain within a safe region that minimizes interference with the retain set.

Rank r	Forget 10% ↑	WMDP Cyber ↓
8	-3.3	30.5
16	-2.8	27.9
32	-1.4	28.4

Table 4: Impact of LoRA rank r on TOFU and WMDP benchmarks.

4.4 Analysis of Hyperparameters

In this section, we investigate the sensitivity of ReGLU to two key hyperparameters: the LoRA rank r and the balance parameter β in Eq. 1. We conduct experiments on TOFU-10% and WMDP-Cyber.

Impact of LoRA Rank r. The rank r determines the dimensionality of the low-rank update ΔW. As shown in Table 4, we evaluate the performance across different ranks. A higher rank provides more degrees of freedom to align the update with the forget-set subspace, potentially leading to more effective unlearning. However, excessively large ranks may increase the risk of interfering with the retain-set knowledge if the orthogonal regularization is not sufficiently strong. Our results indicate that a rank of $r = 16$ or $r = 32$ typically strikes a good balance between unlearning efficacy and utility preservation.

Impact of Balance Parameter β. The parameter $\beta \in [0, 1]$ in Eq. 1 regulates the trade-off between maximizing forget-set variance and minimizing retain-set interference during subspace identification. We vary β while keeping other settings fixed, and report the resulting FQ on TOFU-10% and unlearning performance on WMDP-Cyber. Table 5 shows a clear "sweet spot" at $\beta = 0.3$,

β	Forget 10% ↑	WMDP Cyber ↓
0.1	-2.2	38.2
0.3	-2.1	30.1
0.5	-2.3	36.6
0.7	-3.5	42.6
0.9	-3.6	40.6

Table 5: Impact of the balance parameter β in Eq. 1. Darker colors indicate better performance.

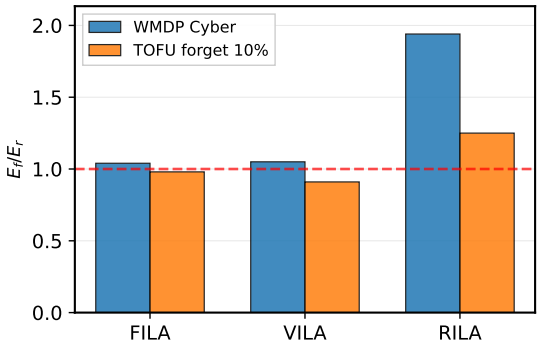

Figure 2: Comparison of activation norms at initialization. RILA (our proposed initialization) achieves a significantly higher forget-to-retain energy ratio compared to FILA and VILA.

which achieves the best trade-off across both benchmarks (highest FQ on TOFU-10% and lowest accuracy on WMDP-Cyber). When β is too small (e.g., 0.1), the objective over-emphasizes maximizing forget-set variance, yielding a less stable subspace for forgetting. Conversely, as β increases (e.g., $\beta \geq 0.7$), the balanced covariance $\mathrm{Cov}_\Delta = (1-\beta)\,\mathrm{Cov}_F - \beta\,\mathrm{Cov}_R$ becomes dominated by the retain term, leading to a subspace that is overly conservative and thus weakens forgetting.

4.5 Analysis of Mechanism

To understand the underlying mechanism of ReGLU, we analyze how the LoRA update interacts with the forget and retain sets at initialization and during training.

Figure 2 shows that RILA achieves a significantly higher forget-to-retain energy ratio at initialization compared to FILA and VILA, indicating that RILA better aligns the LoRA update with the forget-set subspace while minimizing interference with the retain set.

To analyze the subspace of LoRA outputs across different methods, we examine the angular distance between the columns of the LoRA B matrix and the columns of the P_B matrix (detailed in Section 3.2),

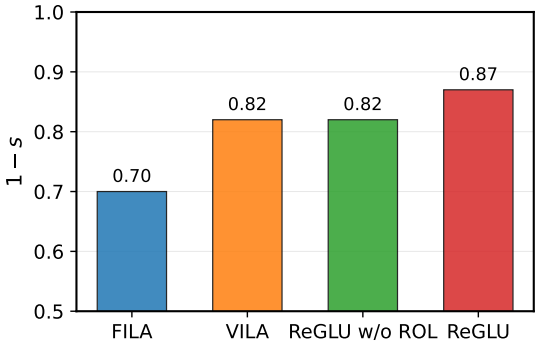

Figure 3: Orthogonality analysis between LoRA B matrix and retain subspace P_B. Higher values of $1-s$ indicate greater orthogonality to the retain representation subspace, which is desirable for effective unlearning while preserving retain-set knowledge.

which measures the influence of LoRA outputs on the retain representation subspace. Specifically, we compute the average pairwise cosine similarity between columns of B and P_B, denoted as $s = \mathrm{avg}_{i,j} \cos^2(B[:,i], P_B[:,j])$, and report $1-s$, which eliminates the effect of the B matrix's scale. Higher values indicate greater orthogonality to the retain subspace, which is desirable for effective unlearning. We conduct experiments on Llama-2-7B with the Forget 5% setting. As shown in Figure 3, ReGLU achieves the highest orthogonality score (0.87), substantially outperforming both FILA (0.70) and VILA (0.82). This demonstrates that our method more effectively maintains LoRA updates orthogonal to the retain representation subspace. Furthermore, removing the ROL component (ReGLU - ROL, 0.82) results in a noticeable decrease in orthogonality, confirming the effectiveness of our subspace-controlled regularization in constraining the LoRA outputs to remain orthogonal to the retain set's principal directions.

4.6 Efficiency Analysis

Table 6 compares the initialization cost on the TOFU benchmark. On Llama2-7B, RILA demonstrates significantly faster initialization compared to FILA across all settings. Compared to VILA, RILA shows comparable efficiency on smaller forget sets, but becomes faster on the Forget 10% setting (0.51 vs. 0.81 GPU hours). This scalability advantage stems from a fundamental difference in computational requirements: while FILA and VILA need to compute gradients to estimate parameter importance maps, RILA only requires a forward pass over the forget set to collect layer outputs,

Method	Params	Rand. SVD	Forget 1%	Forget 5%	Forget 10%
FILA	7B	–	0.71	3.13	23.32
VILA	7B	–	0.09	0.38	0.81
RILA	7B	×	0.50	0.50	0.51
RILA	7B	✓	0.07	0.11	0.12
RILA	70B	✓	0.19	0.27	0.29

Table 6: Efficiency comparison on TOFU benchmark. Time is measured in GPU hours (lower is better). Randomized SVD enables efficient initialization even for 70B-scale models.

followed by covariance computation and eigenvalue decomposition. The most time-consuming operation is the eigenvalue decomposition, whose cost remains nearly constant regardless of dataset size. As a result, RILA's efficiency advantage becomes increasingly pronounced with larger forget sets.

However, the eigenvalue decomposition cost depends on the layer dimension d, which grows with model size. This could become a bottleneck for larger models. Since RILA requires only the top-r eigenvectors, we can leverage randomized SVD to efficiently approximate the leading eigenvectors. Standard eigenvalue decomposition has complexity $O(\min(d^2n, dn^2))$ for an $n \times d$ activation matrix, whereas randomized SVD reduces this to $O(ndr + r^2(n + d))$, where $r \ll \min(n, d)$ is the target rank. For typical LoRA configurations where $r \leq 64$ and $d \sim 4096$, this represents a substantial speedup. As shown in Table 6, with randomized SVD, initialization on Llama2-7B drops to 0.07–0.12 GPU hours. Even on Llama2-70B, RILA maintains efficient initialization at 0.19–0.29 GPU hours, confirming its scalability to larger models.

5 Conclusion

In this work, we introduced ReGLU, a novel LoRA-based method for LLM unlearning. By shifting the focus from parameter importance to representation subspaces, ReGLU effectively addresses the challenges posed by the superposition phenomenon and parameter polysemanticity. Our approach leverages a balanced subspace initialization to align unlearning updates with forget-specific directions and an orthogonal regularization term to protect the principal directions of the retain set. Extensive experiments on the TOFU and WMDP benchmarks demonstrate that ReGLU consistently outperforms state-of-the-art baselines, achieving superior unlearning quality while maintaining high model utility. Our analysis further confirms that ReGLU successfully disentangles forget and retain representations, providing a robust and precise solution for selective forgetting in large language models.

Limitations

Despite its effectiveness, ReGLU has several limitations. First, the method requires computing covariance matrices and performing eigenvalue decomposition for each layer, which introduces a one-time computational overhead during initialization. While this cost is significantly lower than full fine-tuning, it may become non-trivial for extremely large models or high-dimensional representations. Second, the quality of the identified subspaces depends on the representativeness of the small subsets of forget and retain data used for covariance estimation. If these subsets do not accurately capture the underlying distributions, the initialization and regularization may be less effective. Third, our evaluation is primarily focused on the TOFU and WMDP benchmarks. While these are standard in the field, further investigation is needed to assess the generalizability of ReGLU across a broader range of domains and unlearning tasks. Finally, the performance of ReGLU is sensitive to hyperparameters such as the LoRA rank r and the regularization strength λ, which may require careful tuning for different model architectures and datasets.

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

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

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.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. 2025. Towards robust and parameter-efficient knowledge unlearning for llms. In *International Conference on Learning Representations*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

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*.

Ronald A Fisher. 1922. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368.

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)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yejin Kim, Eunwon Kim, Buru Chang, and Junsuk Choe. 2025. Improving fisher information estimation and efficiency for lora-based llm unlearning. *arXiv preprint arXiv:2508.21300*.

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

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *Preprint*, arXiv:2309.05463.

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

Minrui Luo, Fuhang Kuang, Yu Wang, Zirui Liu, and Tianxing He. 2025. Sc-lora: Balancing efficient fine-tuning and knowledge preservation via subspace-constrained lora. *arXiv preprint arXiv:2505.23724*.

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

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072.

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

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Joel A. Tropp. 2012. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. 2025. MiLoRA: Harnessing minor singular components for parameter-efficient LLM

finetuning. 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 4823–4836, Albuquerque, New Mexico. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zeguan Xiao, Siqing Li, Yong Wang, Xuetao Wei, Jian Yang, Yun Chen, and Guanhua Chen. 2026. Modeling llm unlearning as an asymmetric two-task learning problem. *arXiv preprint arXiv:2604.14808*.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Proof of Theorem 1

The proof of Theorem 1 follows the mathematical framework established in SC-LoRA (Luo et al., 2025). We adapt their Theorem 1 to the unlearning problem by reinterpreting the positive task as the forget set and the negative task as the retain set. The key insight—that eigenvectors of the weighted covariance difference capture discriminative directions—applies naturally to our objective of maximizing impact on $\mathcal{D}_f$ while minimizing interference with $\mathcal{D}_r$.

To prove this theorem, we first introduce the concept of orthogonal projection operators.

Definition 1. *Suppose S is a subspace of $\mathbb{R}^n$ of dimension r, and let $\{q_i\}_{i\in[r]}$ be an orthonormal basis of S, then the orthogonal projection operator onto S, denoted Π_S, is defined as:*

$$\Pi_S(x) = \sum_{i=1}^{r} (q_i^\top x) q_i = \sum_{i=1}^{r} (q_i q_i^\top) x. \tag{3}$$

Note: the selection of the orthonormal basis does not affect Π_S.

Proof. We prove this theorem in three steps: (1) establish the relationship between $\|\Delta W x\|_2^2$ and projection onto subspace S; (2) derive the expected projection energy in terms of covariance matrices; and (3) apply Ky Fan's theorem to show that eigenvectors of Cov_Δ maximize the objective.

Step 1: Projection operator representation. Let $\Delta W = BA$. At initialization, $B_{\text{init}} = Q_r$ and $A_{\text{init}} = Q_r^\top W_0$, so $\Delta W = Q_r Q_r^\top W_0$. For any input x, let $h = W_0 x$ be the corresponding output representation. Since Q_r is an orthonormal basis for the subspace S, we have $Q_r Q_r^\top = \Pi_S$. Thus:

$$\Delta W x = Q_r Q_r^\top W_0 x = \Pi_S(h).$$

Therefore, $\|\Delta W x\|_2^2 = \|\Pi_S(h)\|_2^2$.

Step 2: Expected projection energy. Let $\{v_i\}_{i\in[r]}$ be any orthonormal basis that spans S, and denote $\tilde{I}_r = \sum_{i=1}^r v_i v_i^\top$. From the orthonormality of $\{v_i\}_{i\in[r]}$, we have:

$$\begin{aligned}
\tilde{I}_r^\top \tilde{I}_r &= \sum_{i=1}^{r}\sum_{j=1}^{r} v_i v_i^\top v_j v_j^\top \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} v_i \langle v_i, v_j\rangle v_j^\top \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} \delta_{ij} v_i v_j^\top \\
&= \sum_{i=1}^{r} v_i v_i^\top = \tilde{I}_r.
\end{aligned}$$

For either distribution $\mathcal{P} \in \{\mathcal{P}_F, \mathcal{P}_R\}$ with covariance Cov, we have:

$$\begin{aligned}
\mathbb{E}_{h\sim\mathcal{P}}\left[\|\Pi_S(h)\|_2^2\right] &= \mathbb{E}_{h\sim\mathcal{P}}\left[\|\tilde{I}_r h\|_2^2\right] \\
&= \mathbb{E}_{h\sim\mathcal{P}}\left[\mathrm{tr}\left(h^\top \tilde{I}_r^\top \tilde{I}_r h\right)\right] \\
&= \mathbb{E}_{h\sim\mathcal{P}}\left[\mathrm{tr}\left(h^\top \tilde{I}_r h\right)\right] \\
&= \mathbb{E}_{h\sim\mathcal{P}}\left[\mathrm{tr}\left(\tilde{I}_r h h^\top\right)\right] \\
&= \mathrm{tr}\left(\tilde{I}_r \mathbb{E}_{h\sim\mathcal{P}}\left[h h^\top\right]\right) \\
&= \mathrm{tr}\left(\tilde{I}_r\, \mathrm{Cov}\right).
\end{aligned}$$

Substituting into Eq. 1, the reward function becomes:

$$\begin{aligned}
R(S) &= (1-\beta)\mathbb{E}_{h\sim\mathcal{P}_F}\left[\|\Pi_S(h)\|_2^2\right] \\
&\quad - \beta\mathbb{E}_{h\sim\mathcal{P}_R}\left[\|\Pi_S(h)\|_2^2\right] \\
&= (1-\beta)\mathrm{tr}\left(\tilde{I}_r\, \mathrm{Cov}_F\right) - \beta\mathrm{tr}\left(\tilde{I}_r\, \mathrm{Cov}_R\right) \\
&= \mathrm{tr}\left(\tilde{I}_r\, \mathrm{Cov}_\Delta\right).
\end{aligned}$$

Step 3: Optimality via spectral decomposition. Suppose the spectral decomposition of Cov_Δ is $Q\Sigma Q^\top$, where $Q = (q_1, q_2, \ldots, q_{d_{out}})$ is orthogonal and Σ is diagonal with eigenvalues in descending order. Then:

$$\begin{aligned} R(S) &= \mathrm{tr}\left(\tilde{I}_r Q\Sigma Q^\top\right) \\ &= \sum_{i=1}^{r} \mathrm{tr}\left(v_i v_i^\top Q\Sigma Q^\top\right) \\ &= \sum_{i=1}^{r} v_i^\top Q\Sigma Q^\top v_i. \end{aligned}$$

Extend $\{v_i\}_{i\in[r]}$ to a complete orthonormal basis $\{v_i\}_{i=1}^{d_{out}}$ for $\mathbb{R}^{d_{out}}$, and denote $u_i = Q^\top v_i$. Since Q is orthogonal, $\{u_i\}_{i=1}^{d_{out}}$ is also orthonormal. By Ky Fan's theorem,

$$\max_{\{v_i\}_{i\in[r]}} \sum_{i=1}^{r} v_i^\top Q\Sigma Q^\top v_i = \sum_{i=1}^{r} \Sigma_{ii},$$

and this maximum is achieved when $S = \mathrm{span}(\{q_1, q_2, \ldots, q_r\})$, where q_i are the top-r eigenvectors of Cov_Δ. Therefore, initializing $B = Q_r$ and $A = Q_r^\top W_0$ maximizes the objective in Eq. 1.

Uniqueness under eigenvalue gap. If the eigenvalues of Σ satisfy $\lambda_r > \lambda_{r+1}$ (a strict gap), then the maximizing subspace is unique and equals $\mathrm{span}(\{q_1, \ldots, q_r\})$. Let $V = (v_1, \ldots, v_r)$ collect an orthonormal basis of S and define $U = Q^\top V$. Using Q's orthogonality,

$$R(S) = \sum_{i=1}^{r} v_i^\top Q\Sigma Q^\top v_i = \sum_{j=1}^{d_{out}} \Sigma_{jj} \sum_{i=1}^{r} U_{ji}^2.$$

By orthogonality, $0 \le \sum_{i=1}^{r} U_{ji}^2 \le 1$ and $\sum_{j=1}^{d_{out}} \sum_{i=1}^{r} U_{ji}^2 = r$. With a strict spectral gap, the maximum is attained if and only if

$$\sum_{i=1}^{r} U_{ji}^2 = \begin{cases} 1, & 1 \le j \le r, \\ 0, & r+1 \le j \le d_{out}, \end{cases}$$

which is equivalent to $\sum_{i=1}^{r} v_i v_i^\top = \sum_{i=1}^{r} q_i q_i^\top$, hence $S = \mathrm{span}(\{q_1, \ldots, q_r\})$. When $\lambda_r = \lambda_{r+1}$ (no gap), any r-dimensional subspace within the top-eigenspace achieves the same maximum, matching SC-LoRA's discussion. □

B Concentration Bound for Empirical Covariance Estimation

This section provides a concentration bound for empirical covariance estimation, bridging the gap between the population-level covariance matrices used in Theorem 1 and the empirical estimators used in practice. Under a mild bounded-activation assumption, the empirical covariance matrices concentrate sharply in spectral norm, and the balanced covariance used by RILA remains a stable approximation to its population counterpart.

Theorem 2 (Spectral concentration of empirical covariance)**.** *Let $h \in \mathbb{R}^d$ be a random representation vector satisfying $\|h\|_2 \le M$ almost surely, and define its covariance matrix*

$$\Sigma = \mathbb{E}[hh^\top].$$

Given i.i.d. samples $\{h_i\}_{i=1}^N$, let

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} h_i h_i^\top.$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|\hat{\Sigma} - \Sigma\|_2 \le \frac{4M^2 \log(2d/\delta)}{3N} + M^2 \sqrt{\frac{2\log(2d/\delta)}{N}}.$$

In particular, when $N \ge \log(2d/\delta)$, there exists an absolute constant $C > 0$ such that

$$\|\hat{\Sigma} - \Sigma\|_2 \le CM^2 \sqrt{\frac{\log(2d/\delta)}{N}}.$$

Proof. Define

$$X_i = \frac{1}{N}(h_i h_i^\top - \Sigma).$$

Then $\{X_i\}_{i=1}^N$ are independent, zero-mean, symmetric random matrices and

$$\hat{\Sigma} - \Sigma = \sum_{i=1}^{N} X_i.$$

We bound the two quantities required by the matrix Bernstein inequality (Tropp, 2012, Theorem 6.1).

Spectral norm bound. Since $\|h_i h_i^\top\|_2 = \|h_i\|_2^2 \le M^2$ and $\|\Sigma\|_2 \le \mathbb{E}\|h\|_2^2 \le M^2$, we have

$$\|X_i\|_2 \le \frac{\|h_i h_i^\top\|_2 + \|\Sigma\|_2}{N} \le \frac{2M^2}{N}.$$

Hence the Bernstein radius is $R = 2M^2/N$.

Variance bound. Using $\mathbb{E}[hh^\top] = \Sigma$ and $\|h\|_2^2 \le M^2$, we obtain

$$\mathbb{E}[(hh^\top - \Sigma)^2] = \mathbb{E}[\|h\|_2^2 hh^\top] - \Sigma^2 \preceq M^2\Sigma.$$

Therefore,

$$\mathbb{E}[X_i^2] \preceq \frac{M^2\Sigma}{N^2},$$
$$\sigma^2 := \left\|\sum_{i=1}^N \mathbb{E}[X_i^2]\right\|_2 \le \frac{M^2\|\Sigma\|_2}{N} \le \frac{M^4}{N}.$$

Applying matrix Bernstein inequality gives, for any $t > 0$,

$$\Pr\left[\left\|\sum_{i=1}^N X_i\right\|_2 \ge t\right] \le 2d\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

Set $L = \log(2d/\delta)$ and choose

$$t = \frac{LR}{3} + \sqrt{\frac{L^2R^2}{9} + 2L\sigma^2} \le \frac{2LR}{3} + \sqrt{2L\sigma^2}.$$

Then the right-hand side is at most δ, and substituting the bounds on R and σ^2 yields

$$\|\hat{\Sigma} - \Sigma\|_2 \le \frac{4M^2\log(2d/\delta)}{3N} + M^2\sqrt{\frac{2\log(2d/\delta)}{N}}$$

with probability at least $1 - \delta$. When $N \ge \log(2d/\delta)$, the $O(N^{-1})$ term is dominated by the $O(N^{-1/2})$ term up to a universal constant, giving the simplified bound. □

Theorem 3 (Stability of the balanced covariance used by RILA)**.** *Let $\hat{\mathrm{Cov}}_F$ and $\hat{\mathrm{Cov}}_R$ be empirical covariance estimators constructed from N_f forget samples and N_r retain samples, respectively, and define*

$$\hat{\mathrm{Cov}}_\Delta = (1-\beta)\hat{\mathrm{Cov}}_F - \beta\hat{\mathrm{Cov}}_R.$$

Assume the representations from both sets satisfy $\|h\|_2 \le M$ almost surely. Then with probability at least $1 - \delta$,

$$\|\hat{\mathrm{Cov}}_\Delta - \mathrm{Cov}_\Delta\|_2 \le (1-\beta)\epsilon_f + \beta\epsilon_r,$$

where

$$\epsilon_f = \frac{4M^2\log(4d/\delta)}{3N_f} + M^2\sqrt{\frac{2\log(4d/\delta)}{N_f}},$$
$$\epsilon_r = \frac{4M^2\log(4d/\delta)}{3N_r} + M^2\sqrt{\frac{2\log(4d/\delta)}{N_r}}.$$

Consequently, if the eigengap

$$g = \lambda_r(\mathrm{Cov}_\Delta) - \lambda_{r+1}(\mathrm{Cov}_\Delta)$$

is strictly larger than $(1-\beta)\epsilon_f + \beta\epsilon_r$, then the top-$r$ eigenspace recovered from $\hat{\mathrm{Cov}}_\Delta$ is a stable perturbation of the population-optimal subspace, with principal-angle error controlled on the order of $\|\hat{\mathrm{Cov}}_\Delta - \mathrm{Cov}_\Delta\|_2/g$ by standard eigenspace perturbation arguments.

Proof. Apply Theorem 2 to $\hat{\mathrm{Cov}}_F$ and $\hat{\mathrm{Cov}}_R$ separately with failure probability $\delta/2$ each, and take a union bound. On this event, define $\Delta_F = \hat{\mathrm{Cov}}_F - \mathrm{Cov}_F$ and $\Delta_R = \hat{\mathrm{Cov}}_R - \mathrm{Cov}_R$. Then

$$\begin{aligned}\|\hat{\mathrm{Cov}}_\Delta - \mathrm{Cov}_\Delta\|_2 &= \|(1-\beta)\Delta_F - \beta\Delta_R\|_2 \\ &\le (1-\beta)\|\Delta_F\|_2 + \beta\|\Delta_R\|_2 \\ &\le (1-\beta)\epsilon_f + \beta\epsilon_r.\end{aligned}$$

The eigenspace stability statement then follows from standard perturbation bounds for symmetric matrices: once the perturbation magnitude is smaller than the spectral gap g, the leading eigenspace of $\hat{\mathrm{Cov}}_\Delta$ remains close to that of Cov_Δ. □

C Algorithm

We provide the complete algorithmic description of ReGLU in Algorithm 1.

D Implementation and Training Details

D.1 Hardware and Environment

Unless otherwise specified, all experiments were conducted on a single NVIDIA L20 GPU (48GB VRAM). Experiments on WMDP were conducted on a single NVIDIA A100 GPU (40GB VRAM). Our implementation is based on PyTorch 2.5.1 (CUDA 12.1) and the Hugging Face ecosystem, including Transformers 5.0.0.dev0, Tokenizers 0.22.1, and PEFT 0.17.1. We used DeepSpeed with ZeRO Stage 2 for memory optimization and launched runs via `torchrun`. All experiments were trained using BF16 precision by default.

D.2 Models and LoRA Configurations

We report results on three backbone models depending on the benchmark: Llama2-7B and Phi-1.5B for TOFU, and Zephyr-7B-β for WMDP. For parameter-efficient updates, we apply LoRA adapters to the following linear projections in every transformer block: { `q_proj`, `k_proj`, `v_proj`,

Algorithm 1 Representation Subspace-Controlled Unlearning (ReGLU)

Require: Pre-trained model $\mathcal{M}$ with parameters $\boldsymbol{\theta}$, forget set $\mathcal{D}_f$, retain set $\mathcal{D}_r$, LoRA rank r, retain subspace dimension k, hyperparameters β, γ, λ

Phase 1: Initialization
Sample N_f examples from $\mathcal{D}_f$ and N_r examples from $\mathcal{D}_r$
for all trainable parameters in $\mathcal{M}$ **do**
    Feed forget samples, collect outputs $H_f = [h_1^{(f)}, h_2^{(f)}, \ldots, h_{N_f}^{(f)}]^\top \in \mathbb{R}^{N_f \times d}$
    Feed retain samples, collect outputs $H_r = [h_1^{(r)}, h_2^{(r)}, \ldots, h_{N_r}^{(r)}]^\top \in \mathbb{R}^{N_r \times d}$
    Compute $\mathrm{Cov}_F \leftarrow \frac{1}{N_f} H_f^\top H_f$
    Compute $\mathrm{Cov}_R \leftarrow \frac{1}{N_r} H_r^\top H_r$
    Compute $\mathrm{Cov}_\Delta \leftarrow (1-\beta)\,\mathrm{Cov}_F - \beta\,\mathrm{Cov}_R$
    Perform eigenvalue decomposition on Cov_Δ
    Extract top-r eigenvectors $Q_r = (q_1, q_2, \ldots, q_r)$
    Initialize $B_{\text{init}} \leftarrow Q_r$
    Initialize $A_{\text{init}} \leftarrow Q_r^\top W_0$
    Compute $W_{\text{res}} \leftarrow W_0 - B_{\text{init}} A_{\text{init}}$ ▷ Residual weight; frozen during training
    Perform eigenvalue decomposition on Cov_R
    Extract top-k eigenvectors $P_B = (p_1, p_2, \ldots, p_k)$
end for
Phase 2: Training
for each training iteration **do**
    Sample mini-batch from $\mathcal{D}_f$ and $\mathcal{D}_r$
    Compute forget loss: $\mathcal{L}_{\text{forget}} \leftarrow \mathcal{L}_{\text{IHL}}(\mathcal{D}_f)$ or $\mathcal{L}_{\text{GA}}(\mathcal{D}_f)$
    Compute retain loss: $\mathcal{L}_{\text{retain}} \leftarrow \mathcal{L}_{\text{CE}}(\mathcal{D}_r)$
    Compute orthogonal loss: $\mathcal{L}_{ROL} \leftarrow \|B^\top P_B\|_F^2$
    Compute total loss: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{forget}} + \gamma \mathcal{L}_{\text{retain}} + \lambda \mathcal{L}_{ROL}$
    Update LoRA parameters $\{A, B\}$ via gradient descent
end for
return Unlearned model $\mathcal{M}'$ with updated LoRA adapters

o_proj, gate_proj, up_proj, down_proj }. Unless otherwise stated, we use LoRA rank $r = 32$ and set the scaling factor to $\alpha = 2r$ (thus $\alpha = 64$). The LoRA dropout is set to 0.0 for TOFU and 0.05 for WMDP.

D.3 Hyperparameters Search Space

We performed a grid sweep over key hyperparameters. Across all experiments, we fix the retain loss weight $\gamma = 1.0$ and the retain subspace dimension $k = 128$.

TOFU. All TOFU runs were trained for 5 epochs with batch size 4 and gradient accumulation steps 8 (effective batch size 32), using weight decay 0.01. For Llama2-7B, we swept the learning rate over $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$, and swept $\lambda \in \{0.5, 0.7\}$ and $\beta \in \{0.5, 0.7\}$. For Phi-1.5B, we used the same grid, additionally including 2×10^{-4} in the learning-rate sweep, i.e., $\{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}\}$, and swept $\lambda \in \{0.5, 0.7\}$ and $\beta \in \{0.5, 0.7\}$.

WMDP. We swept the learning rate over $\{1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$ and $\lambda \in \{0.1, 0.5, 0.7\}$, with max_steps=100. All WMDP experiments used LoRA rank $r = 32$ and $\alpha = 64$.

D.4 Model Selection Criteria

For TOFU, we selected hyperparameters that yield a favorable trade-off between Forget Quality (FQ) and Model Utility (MU). For WMDP, we enforced a utility constraint based on MMLU, requiring MMLU to be at least 95% of the original model's performance. Among feasible configurations, we selected those that minimize accuracy on the target corpora (bio/cyber) within the swept hyperparameter grid.

D.5 Best Configurations Reported for WMDP

For reproducibility, we report the exact best-performing configurations used in our main WMDP results. All configurations below are evaluated on checkpoints of steps=100.

- **GD + ReGLU (bio):** $r = 32$, $\alpha = 64$, lr $= 1 \times 10^{-5}$, $\beta = 0.7$, $\lambda = 0.5$.
- **GD + ReGLU (cyber):** $r = 32$, $\alpha = 64$, lr $= 3 \times 10^{-5}$, $\beta = 0.5$, $\lambda = 0.5$.
- **IHL + ReGLU (bio):** $r = 32$, $\alpha = 64$, lr $= 5 \times 10^{-5}$, $\beta = 0.5$, $\lambda = 0.1$.
- **IHL + ReGLU (cyber):** $r = 32$, $\alpha = 64$, lr $= 3 \times 10^{-5}$, $\beta = 0.5$, $\lambda = 0.5$.