Invariance Makes LLM Unlearning Resilient Even to Unanticipated Downstream Fine-Tuning

Changsheng Wang[1], Yihua Zhang[1], Jinghan Jia[1], Parikshit Ram[2], Dennis Wei[2], Yuguang Yao[1], Soumyadeep Pal[1], Nathalie Baracaldo[2], Sijia Liu[1,2]

[1]Michigan State University, [2]IBM Research

Paper Code

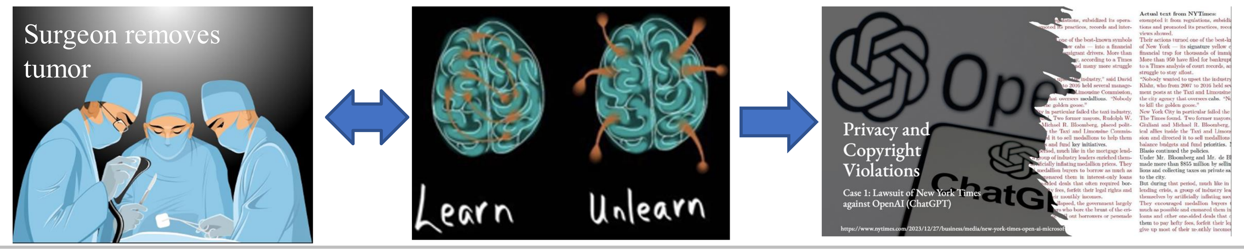
➢ What Is LLM Unlearning?

- LLM unlearning aims to remove *undesirable* learned information from a trained model, while preserving overall utility[1].

$$\boldsymbol{\theta}_{\mathrm{u}} = \arg\min_{\boldsymbol{\theta}} \underbrace{\ell_{\mathrm{f}}(\boldsymbol{\theta};\, \mathcal{D}_{\mathrm{f}})}_{\text{Forget}} + \lambda \underbrace{\ell_{\mathrm{r}}(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{r}})}_{\text{Retain}}$$

Here, $\mathcal{D}_{\mathrm{f}}$ is the forget set to be unlearned, and $\mathcal{D}_{\mathrm{r}}$ is the retain set to preserve utility.

Unlearning: A surgery to AI   Application of unlearning



➢ Unlearning Vulnerability in the Face of Downstream Fine-tuning

- Knowledge removed through unlearning can be rapidly recovered via post-unlearning fine-tuning, even when the new data is unrelated[2].

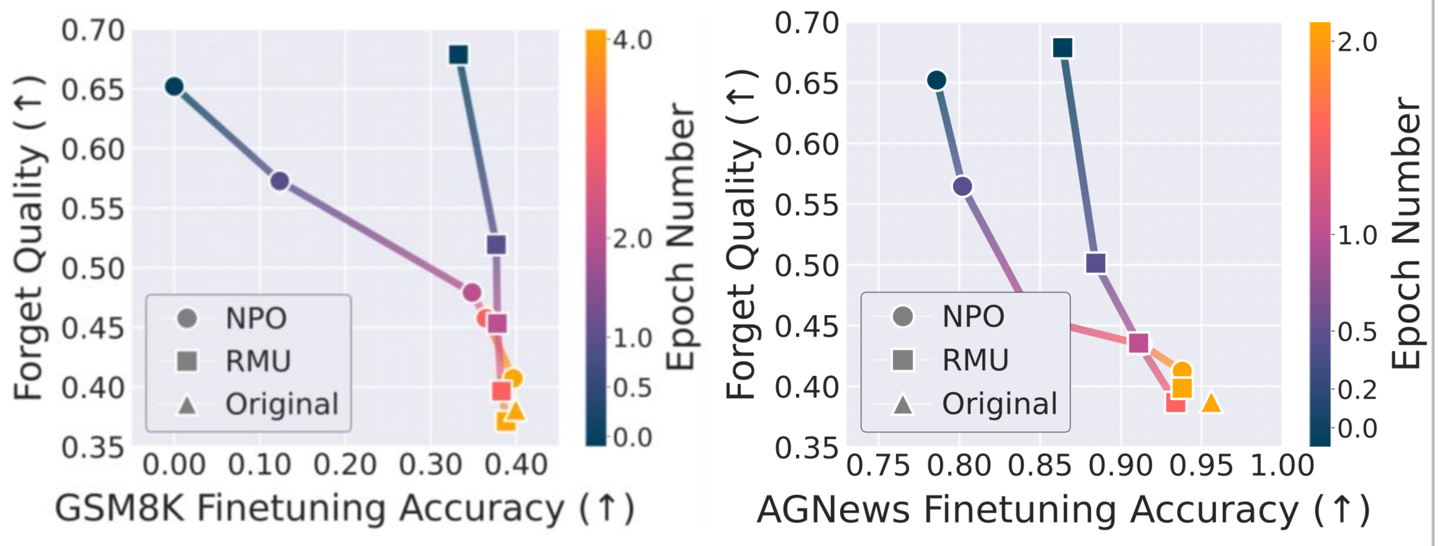


Figure 1. Motivating example: Fine-tuning breaks existing unlearning methods (NPO and RMU) on the WMDP using Zephyr-7B-beta [3]. Forgetting is measured by 1 - WMDP accuracy. Color indicates the fine-tuning epochs, from no tuning to the point where performance matches that of full fine-tuning ('Original').

➢ IRM Principle: Learning Invariant Predictor Across Environments

- **Invariant Risk Minimization (IRM)** [4] aims to learn a model that remains simultaneously optimal across all training environments. A tractable formulation is known as IRMv1 [4], formulated as:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \underbrace{\ell_{\mathrm{ERM}}(\boldsymbol{\theta})}_{\text{ERM}} + \underbrace{\lambda \sum_{i=1}^{N} \left\| \nabla_{\boldsymbol{w}|\boldsymbol{w}=1} \ell_i(\boldsymbol{w} \circ \boldsymbol{\phi};\, \mathcal{D}_i) \right\|}_{\text{Invirance Regularization}}$$

Here, $\boldsymbol{w}$ is invariant predictor, $\boldsymbol{\phi}$ is shared representation network, the composition $\boldsymbol{\theta} = \boldsymbol{w} \circ \boldsymbol{\phi}$ yields the full model, N is the number of training environments, and $\mathcal{D}_i$ is the dataset for the i-th environment. By IRMv1, $\boldsymbol{w} = 1$ can be regarded as a virtual (scalar) predictor such that $\boldsymbol{\theta} = \boldsymbol{\phi}$.

- **Insight:** This IRM mechanism, originally designed for improving domain generalization, inspires us to promote the invariance of unlearning against additional fine-tuning on the unlearned model.

➢ Invariant LLM Unlearning (ILU)

- We adapt IRM to unlearning by replacing the ERM loss with an unlearning objective ℓ_u, while keeping the invariance regularization to resist downstream fine-tuning

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \; \ell_u(\boldsymbol{\theta}) + \lambda \sum_{i=1}^{N} \left\| \nabla_{\boldsymbol{w}|\boldsymbol{w}=1} \ell_i(\boldsymbol{w} \circ \boldsymbol{\phi};\, \mathcal{D}_i) \right\|$$

Here, $\mathcal{D}_i$ encodes the fine-tuning environment (e.g., GSM8K or AGNews), unrelated to unlearning.

- The invariance regularization encourages $\boldsymbol{\theta}$ to be robust to fine-tuning across all $\mathcal{D}_i$.

➢ Analysis via Task Vector

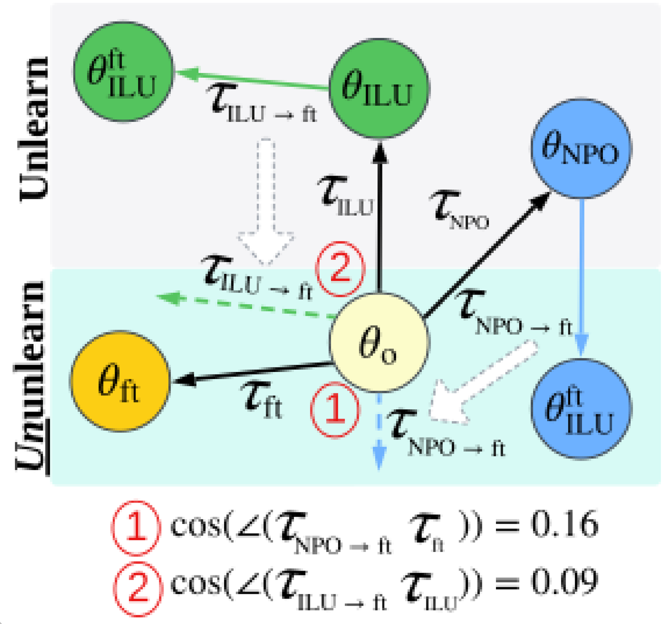


Figure 2. Illustration of ILU's improved unlearning robustness compared to NPO through the relationship between unlearning task vector and fine-tuning task vector on WMDP with Zephyr-7b-beta.

➢ Single Fine-tune Set Suffices for ILU

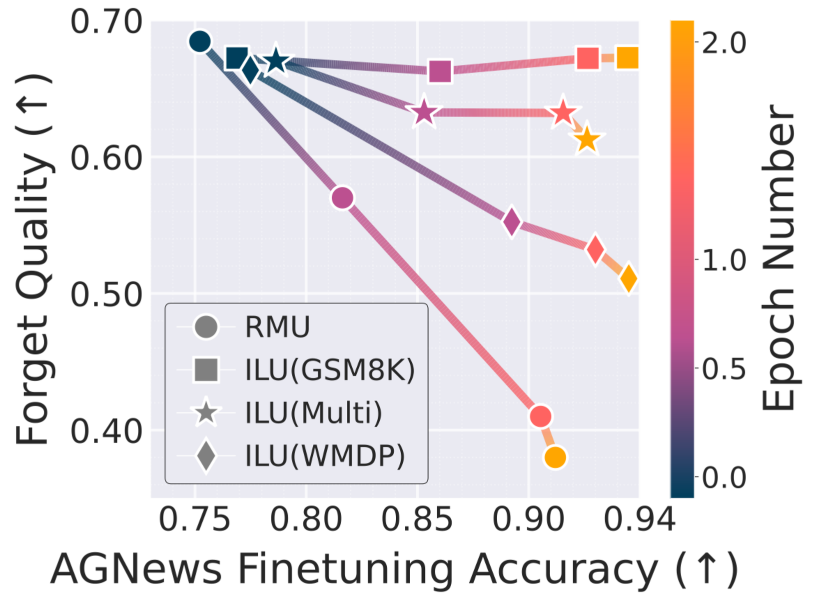


Figure 3. A single fine-tuning dataset suffices for preserving unlearning efficacy against fine-tuning. Here, ILU(Multi) adopts GSM8K, AGNews, and WinoGrande as multiple invariance sources in regularization.

➢ Experiment Results Highlights

- **Effectiveness of ILU on WMDP Dataset**

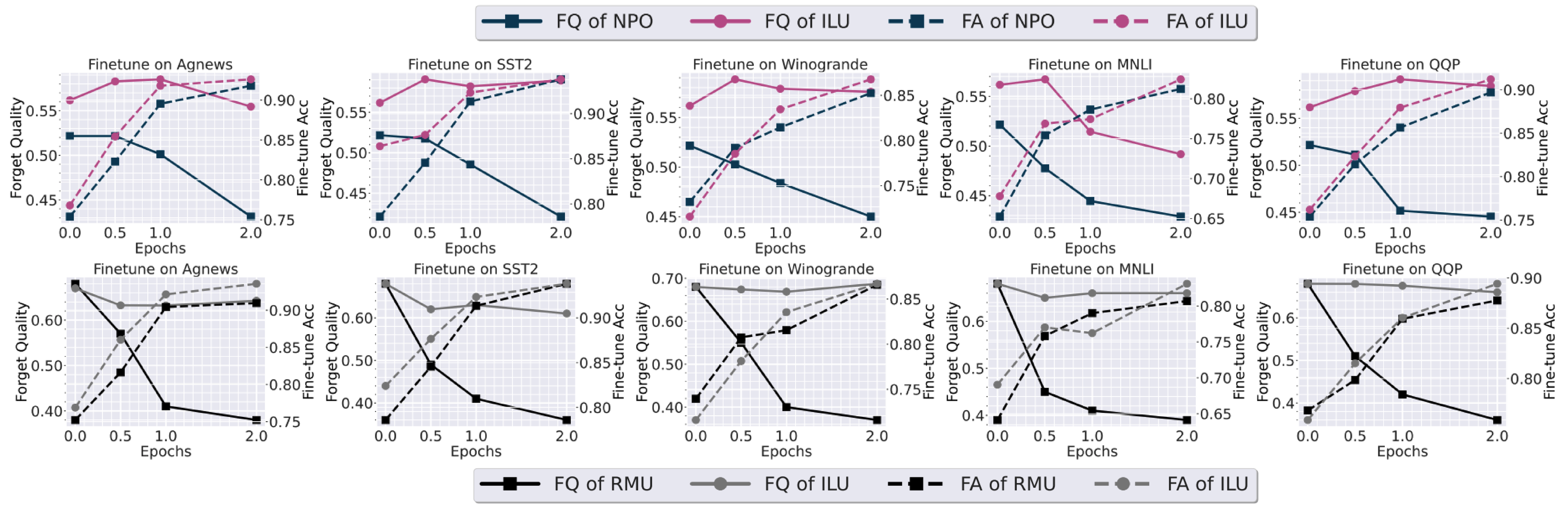


Figure 3. Resilience of unlearning to downstream fine-tuning across different fine-tuning epochs. Each subplot represents a downstream fine-tuning dataset. The x-axis denotes the fine-tuning epoch, with the maximum number set to ensure convergence and satisfactory fine-tuning performance for each downstream task.

- **ILU on MUSE Dataset**

Method	MUSE-News VerbMem on $\mathcal{D}_{\mathrm{f}}$ ↓	MUSE-News KnowMem on $\mathcal{D}_{\mathrm{f}}$ ↓	MUSE-News KnowMem on $\mathcal{D}_{\mathrm{r}}$ ↑	MUSE-News FA ↑	MUSE-Books VerbMem on $\mathcal{D}_{\mathrm{f}}$ ↓	MUSE-Books KnowMem on $\mathcal{D}_{\mathrm{f}}$ ↓	MUSE-Books KnowMem on $\mathcal{D}_{\mathrm{r}}$ ↑	MUSE-Books FA ↑
Original model	58.40	63.90	55.20	-	99.80	59.40	66.90	-
Pre-Finetune								
NPO	2.53	**40.76**	36.25	-	**0.00**	**0.00**	**57.19**	-
+ILU(GSM8K)	**0.00**	46.97	**41.90**	-	**0.00**	**0.00**	45.20	-
Post-Finetune on GSM8K								
NPO	35.38	52.73	**47.29**	16.53	9.69	38.03	**63.29**	5.84
+ILU(GSM8K)	**0.46**	**49.97**	42.90	**18.64**	**0.00**	**31.47**	56.30	**6.08**
Post-Finetune on AGNews								
NPO	13.96	53.87	44.43	**94.20**	1.39	36.35	**66.00**	**94.00**
+ILU(GSM8K)	**0.00**	**44.95**	**44.97**	94.00	**0.00**	**14.37**	61.17	93.80
Post-Finetune on SST-2								
NPO	3.63	**44.12**	**38.83**	**97.20**	1.61	31.88	**63.17**	96.80
+ILU(GSM8K)	**0.00**	**44.12**	36.18	97.00	**0.00**	**23.63**	60.62	**97.00**
Post-Finetune on WinoGrande								
NPO	57.27	64.96	**54.36**	**67.40**	2.86	38.00	**66.67**	**60.22**
+ILU(GSM8K)	**0.00**	**48.68**	44.58	59.00	**0.00**	**20.03**	61.34	59.27
Post-Finetune on MNLI								
NPO	32.54	48.61	**46.54**	**85.20**	8.58	33.42	**62.84**	81.56
+ILU(GSM8K)	**0.00**	**47.84**	45.65	84.46	**0.00**	**28.54**	61.32	**83.68**
Post-Finetune on QQP								
NPO	33.46	54.21	45.86	**93.00**	9.57	31.58	**66.10**	91.68
+ILU(GSM8K)	**2.07**	**46.17**	**47.68**	92.86	**0.00**	**24.78**	63.54	**92.80**

Table 1. Comparison of ILU and NPO on MUSE-News and MUSE-Books benchmarks, evaluating performance both before and after fine-tuning.

- **Generalization of ILU**

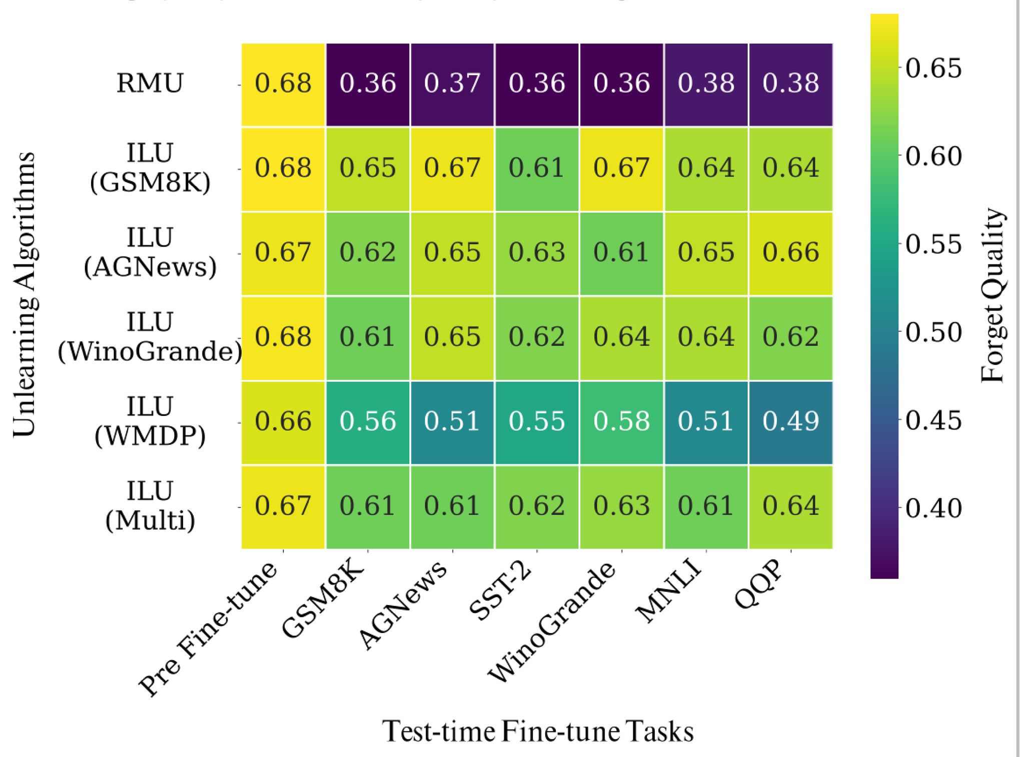


Figure 4. Generalization of ILU to unseen fine-tuning tasks during evaluation. A heatmap of forget quality on WMDP is presented for RMU and its ILU variants, demonstrating unlearning robustness under various unlearning training and downstream fine-tuning settings. Each row corresponds to an unlearning approach, and each column represents a post-unlearning fine-tuning setting.

[1] Liu, Sijia, et al. "Rethinking machine unlearning for large language models." Nature Machine Intelligence (2025): 1-14.

[2] Hu, Shengyuan, et al. "Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning." ICLR2025.

[3] Li, Nathaniel, et al. "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning." ICML 2024.

[4] Arjovsky, Martin, et al. "Invariant risk minimization." arXiv preprint arXiv:1907.02893 (2019).

Contact: {wangc168, liusiji5}@msu.edu