Model Immunization by Trapping Harmful Finetuning

Najibul Haque Sarker[†]
najibulhaque@vt.edu

Zaber Ibn Abdul Hakim[†]
zaberhakim666@vt.edu

Alvi Md Ishmam[†]
alvi@vt.edu

Chia-Wei Tang[†]
cwtang@vt.edu

Chris Thomas[†]
christhomas@vt.edu

[†]Virginia Tech

Abstract

Model immunization is a new technique of protecting models against downstream harmful fine-tuning while remaining useful on intended tasks. Prior works utilize condition number based regularizers to ill-condition the optimization landscape for harmful tasks. However, the induced protection does not guarantee that immunization will persist. In this work, we introduce the novel concept of creating a trap in the landscape, so that harmful finetuning optimization will be trapped in an unoptimized minima. We propose a geometry-aware trap-inducing objective, which limits multi-step harmful loss reduction to the expected local geometry-based loss. Furthermore, to properly evaluate immunization retainment, we introduce an extrinsic metric, Relative Fine-Tuning Deviation (RFD). Across multiple pretrained backbones and datasets, we show our method increases resistance to harmful adaptation and preserves primary-task accuracy, outperforming curvature-only baselines on RFD while remaining competitive on standard utility metrics.

1 Introduction

Open model release has unlocked extraordinary downstream utility but also enables rapid adaptation to undesirable tasks. The model immunization task aims to tackle this aspect by studying parameters that remain useful on intended tasks yet resist (or significantly raise the cost of) adaptation on restricted domains [Zheng and Yeh, 2024, Deng et al., 2024, Zheng et al., Huang, 2025]. Prior work has framed this goal as task blocking or non–fine-tunable learning. The model is either conditioned during pretraining or in a separate post-training phase so that subsequent fine-tuning on harmful tasks is hampered, thus increasing the cost of undesirable downstream adaptation while preserving performance on desired tasks.

Current works for model immunization have largely treated the problem in the meta-learning context [Deng et al., 2024, Zheng and Yeh, 2024, Zheng et al., 2024]. These methods simulate an adversary during training and optimize an initialization that resists subsequent adaptation on restricted domains. However, as these methods are learning a bad initialization for the harmful task, the protection is sensitive to the specific setting of the meta-learning optimization and may not transfer when the attacker deviates from these choices.

A complementary line of work shifts focus from meta-learning to conditioning the model's geometry [Zheng et al.]. Instead of simulating fine-tuning, it directly shapes curvature by controlling the Hessian's condition number, which is the ratio between extreme singular values that governs how easily optimization can progress. By reducing curvature (well-conditioning) of the model's loss landscape in desired directions and increasing it (ill-conditioning) for restricted ones, they hinder the ease of harmful adaptation intrinsically. More related works can be found in Appendix A.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Lock-LLM: Prevent Unauthorized Knowledge Use from LLMs.

While curvature control can hinder harmful fine-tuning, it only affects current parameters and not the overall optimization path or destination. Condition number regularization can be optimized by gradient descent and thus may facilitate adversarial search for descent directions that still yield large loss decreases on restricted domains. Thus, curvature alone cannot ensure negligible harmful adaptation across multiple fine-tuning steps. What is needed is a criterion that also targets the destination of the optimization, and conditions it to be a suboptimal local minima. We refer to such spaces as **traps** for the optimization process.

We propose constructing such *traps* in the optimization path, so that even after plausible fine-tuning steps, the expected loss reduction on the restricted domain is minimal. Intuitively, a trap is a region where adaptation looks promising locally yet yields little global improvement, making it inefficient for an adversary to escape without essentially re-training. Concretely, we simulate the actual multi-step loss reduction under adaptation and reduce it to the estimated expected loss reduction from the initial parameter setting, thus trapping the optimization by local geometry. This setting complements curvature-based conditioning, which slows down optimization, while traps restrict the destination of the optimization into a local optima.

We validate the proposed methodology empirically on two widely used pretrained models and three separate downstream datasets. Existing methods use `RIR` (Relative Immunization Ratio) metric [Zheng et al.], which is an intrinsic measure of immunization and can be unreliable. We introduce a new metric `RFD` (Relative Finetuning Deviation), which is an extrinsic immunization criterion and overcomes previous issues. We observe considerable improvement in model immunity across all the datasets and backbones, conveying the effectiveness of our proposed methodology.

The main contributions of our work are as follows: **(1)** we introduce a trap-based immunization objective that minimizes expected loss reduction on restricted domains. In combination with curvature-based regularizers that shape the local geometry, these traps suppress multi-step progress in a local optima; **(2)** we propose RFD, an extrinsic and reliable metric that captures persistent divergence from baseline adaptation in linear probing, thereby addressing the limitations of intrinsic conditioning ratio metrics; and **(3)** we provide comprehensive experiments on ImageNet-pretrained backbones and multiple restricted domains, showing consistent gains in immunization capability while maintaining strong pretraining performance.

2 Preliminaries

This section provides the setting of model immunization, the background of condition-number based methods and limitations of pure curvature conditioning.

Model immunization setting. We adopt the model immunization framework [Zheng and Yeh, 2024] in which a model is trained to be simultaneously useful on a primary task while being resistant to harmful fine-tuning. Concretely, we assume access to two datasets: a primary dataset D_P for model pretraining, where high performance should be retained, and a harmful dataset D_H, where fine-tuning should be intentionally impeded. Following prior work on condition-number based immunization [Zheng et al.], we consider a feature extractor f_θ parameterized by θ, coupled with a lightweight linear head ω, that is already pretrained with the primary dataset D_P. An adversary is modeled as using transfer learning via linear probing [Zhuang et al., 2020], by attaching a new linear head ω_H to f_θ and optimizing on D_H. In the adversary setting, f_θ is frozen and only ω_H is learnable. The goal of immunization is to convert f_θ to an immunized version f_θ^I such that optimization on D_H using $f_\theta^I(D_H) \cdot \omega_H$ is impeded, while performance on D_P using $f_\theta^I(D_P) \cdot \omega$ remains the original.

Condition number–based model immunization. A central quantity in this setting is the *condition number* [Gloub and Van Loan, 1996] of a matrix S,

$$\kappa(S) = \|S\|_2 \cdot \|S^\dagger\|_2 = \frac{\sigma_S^{max}}{\sigma_S^{min}}, \tag{1}$$

where $S^\dagger$ denotes the pseudoinverse, and the max and min singular values are denoted by σ_S^{max} and σ_S^{min}. The condition number captures the spread of singular values, and thus the curvature of the loss landscape when S is the Hessian H matrix. Multiple works have looked at the implications of the condition number κ for gradient-based optimization [Nenov et al., 2024, Boyd and Vandenberghe, 2004], where lower κ creates a smooth gradient landscape which indicates faster convergence during fine-tuning (well-conditioned), and higher κ indicates slower fine-tuning performance due to the landscape having high curvature (ill-conditioned). Nenov et al. [2024] introduced a regularizer that

minimizes κ, which was adapted into two regularizers by Zheng et al. for the purpose of inducing ill-conditioned curvature for optimization on D_H, while maintaining D_P optimization:

$$R_{well}(S) = \frac{1}{2}\|S\|_2^2 - \frac{1}{2\nu}\|S\|_{\sigma_S^{max}}^2 \tag{2}$$

$$R_{ill}(S) = \frac{1}{\frac{1}{2}\|S\|_2^2 - \frac{1}{2\nu}\|S\|_{\sigma_S^{min}}^2} \tag{3}$$

By utilizing $R_{\text{ill}}(H_H)$ and $R_{\text{well}}(H_P)$ as regularizers, where H_H and H_P stands respectively for the Hessian for D_H and D_P optimization, κ for the harmful task can be increased while κ for the primary task decreases. Zheng et al. showed theoretically that immunization to harmful fine-tuning can be characterized using this formulation in a linear-model setting, and extended to show effectiveness of the method to non-linear models empirically. Here, we only focus on the non-linear model setting, and the effect of curvature-based conditioning on model-immunization.

Limitations of Pure Curvature Conditioning. While curvature shaping is effective, relying solely on κ-based regularization is insufficient for robust immunization. A large condition number slows down harmful fine-tuning but does not prevent eventual convergence if the adversary has sufficient compute or employs alternative optimization strategies, such as using a higher learning rate. Moreover, condition numbers reflect only spectral properties of the loss Hessian and provide no guarantees on where optimization trajectories lead. As a result, models can still exhibit harmful loss reduction even in high–κ regimes. This motivates extending beyond curvature to explicitly design deceptive landscapes or **traps** in which harmful fine-tuning progress is intrinsically limited.

3 Methodology

In this section, we first introduce our core contribution which is a trap-inducing loss for harmful updates. Then we formulate the final immunization objective, and discuss suitable evaluation metrics.

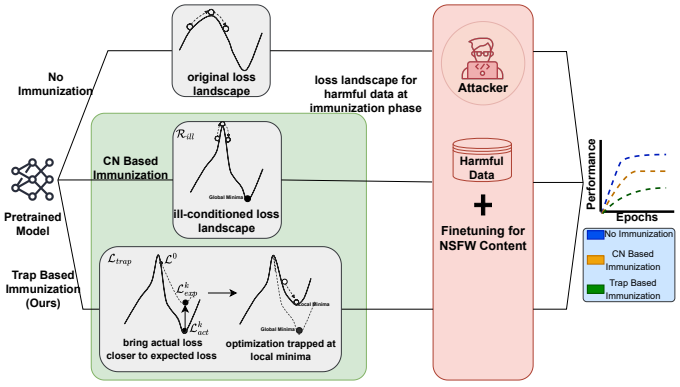

Figure 1: Comparison between harmful-task loss landscapes for non-immunized model, condition number (CN) based immunization [Zheng et al.], and our introduced trap inducing loss based immunization. For the non-immunized model, the curvature is smooth and gradient descent easily finds the global optima. For the CN-based immunization, using Equation 2 introduces high curvature for the harmful task, but it can also lead a path to a better optima as well. Our formulation (Equation 6) minimizes the difference between the actual multi-step loss $\mathcal{L}_{act}^k$ and the expected loss $\mathcal{L}_{exp}^k$, which induces deceptive plateaus / slow-descent regions. Our immunized model is the most resistant to harmful finetuning as shown in downstream performance.

3.1 Trap Inducing Loss

We introduce trap-inducing loss, a geometry-aware regularizer that explicitly limits the extent to which harmful fine-tuning can reduce loss. The key idea is to construct local regions of the parameter space where optimization on the harmful dataset is deceptive. Although gradients may initially appear promising, the actual loss improvement is no better (and often strictly worse) than what local curvature predicts. As a result, harmful fine-tuning is steered into shallow basins from which substantial progress is difficult.

Let L_H denote the harmful-task loss for the harmful dataset D_H, and let θ^0 be the current parameters with $g_0 = \nabla L_H(\theta^0)$ and $H_0 = \nabla^2 L_H(\theta^0)$, and θ^k are future parameters after k optimization steps. Here, we consider $\Delta\theta \approx \theta^k$ - θ^0.

Taylor expansion and approximation. For a small update $\Delta\theta$, a second-order Taylor expansion around θ^0 gives

$$L_H(\theta^k) \approx L_H(\theta^0 + \Delta\theta) = L_H(\theta^0) + g_0^\top \Delta\theta + \tfrac{1}{2}\,\Delta\theta^\top H_0\,\Delta\theta + R_3$$

where R_3 denotes higher-order terms. In practice, these higher-order contributions are negligible for sufficiently small steps, so the quadratic approximation serves as the local model of the loss landscape.

Expected loss reduction. The quadratic surrogate predicts a reduction of

$$\Delta L_{\text{exp}} = L_H(\theta^0) - L_H(\theta^0 + \Delta\theta) = -(g_0^\top \Delta\theta + \tfrac{1}{2}\,\Delta\theta^\top H_0\,\Delta\theta) \tag{4}$$

This expression captures the improvement the local curvature permits, a gain that is moderated by the curvature penalty along the descent direction.

Actual loss reduction. The actual realized decrease after k steps of optimization is

$$\Delta L_{\text{act}} = L_H(\theta^0) - L_H(\theta^k) \tag{5}$$

If ΔL_{act} substantially exceeds ΔL_{exp}, then the landscape around θ^0 provided an unexpectedly favorable descent path, enabling harmful fine-tuning to progress faster than the local model predicts. Here, the Taylor series approximation is based on a small $\Delta\theta$ change, but we are taking k steps of optimization. So we expect, $\theta^k - \theta^0 \gtrsim \Delta\theta$ and $\Delta L_{\text{act}} > \Delta L_{\text{exp}}$.

Trap inducing loss formulation. We introduce the **trap inducing loss**, which penalizes any surplus improvement during model optimization beyond the initial expectation:

$$\mathcal{L}_{\text{trap}}(\theta^0) = \text{softplus}\Big(\Delta L_{\text{act}} - \Delta L_{\text{exp}}\Big) \tag{6}$$

This formulation enforces a deceptive landscape for the harmful-task loss L_H. Gradients may initially signal meaningful descent directions. However, the landscape is sculpted so that the realized loss decrease is bounded by the local quadratic model. Fine-tuning thus becomes trapped in a local minimum. Over multiple steps, this results in attractor basins in which harmful optimization requires disproportionately many updates to escape.

Synergy with condition number regularization. Condition-number based regularizers slow harmful adaptation by increasing ill-conditioning, but they do not guarantee that the harmful loss will not eventually decrease if sufficient optimization effort is applied. The trap loss complements this approach by bounding the realized decrease relative to its quadratic prediction, ensuring that even when ill-conditioned directions are eventually navigated, the actual improvement remains limited. In combination, condition-number and the trap loss yield models that are both spectrally ill-conditioned and geometrically deceptive, providing substantially stronger immunization than either method alone.

3.2 Model Immunization Objective

We incorporate trap inducing loss with the condition number based curvature conditioning for our final model immunization process. Given a pretrained feature extractor θ, pretrained linear head ω, primary dataset D_P, and harmful dataset D_H, our final objective is:

$$\min_{\theta} \mathcal{L}_{\text{trap}}(\theta) + R_{ill}(H_H) + R_{well}(H_P) + \mathcal{L}(D_P, \omega, \theta) \tag{7}$$

where $\mathcal{L}_{\text{trap}}$ denotes our trap inducing loss, R_{ill} and R_{well} denotes regularizers to maximize and minimize condition numbers of Hessian H_H for D_H optimization and Hessian H_P for D_P optimization respectively, and finally $\mathcal{L}(D_P, \omega, \theta)$ refers to the original primary dataset optimization objective. The output of this optimization will be an immunized feature extractor θ_I.

3.3 Evaluation Metric

To evaluate immunization, Zheng et al. introduced the **Relative Immunization Ratio (RIR)** metric, which measures the ratio of condition number of Hessian:

$$\texttt{RIR} \triangleq \left(\frac{\kappa(H_H(\theta_I))}{\kappa(H_H(\theta_0))}\right) \Big/ \left(\frac{\kappa(H_P(\theta_I))}{\kappa(H_P(\theta_0))}\right) \tag{8}$$

Table 1: Quantitative results of immunization of ImageNet pretrained ResNet18 and ViT models over 3 finetuning datasets. We report both RIR and RFD to show immunization quality, while test accuracy onthe primary dataset ImageNet D_P is shown to ensure original utility retainment.

D_H	Method	ResNet18			ViT		
		RIR ↑	D_P Test Acc. (%) ↑	RFD ↑	RIR ↑	D_P Test Acc. (%) ↑	RFD ↑
	Init. θ_0	1.0	67.04	-	1.0	82.37	-
Cars	R_{ill} *Only*	1.078	63.58	2.28	4.578	82.01	2.93
	IMMA	1.002	63.71	2.36	0.781	80.95	7.57
	Opt κ	1.007	63.73	0.16	3.47	**82.68**	7.57
	CN	3.521	62.27	10.06	**84.155**	82.01	33.32
	Ours	**43.920**	**65.99**	**47.19**	70.128	82.48	**38.77**
Food101	R_{ill} *Only*	1.239	63.8	3.333	1.358	80.14	2.29
	IMMA	1.021	**65.41**	0.22	0.799	81.27	1.38
	Opt κ	1.082	63.80	0.37	42.49	82.36	5.08
	CN	27.47	58.04	4.62	**96.300**	82.20	**8.38**
	Ours	**64.845**	64.54	**19.04**	86.203	**82.69**	7.46
Country211	R_{ill} *Only*	1.126	66.15	6.50	1.105	76.37	4.78
	IMMA	1.002	**67.33**	1.03	0.796	81.59	7.80
	Opt κ	1.007	67.31	0.68	7.183	82.51	3.35
	CN	**25.701**	61.92	21.16	71.41	**82.97**	26.44
	Ours	25.04	60.16	**28.32**	**86.303**	82.91	**28.35**

While informative, RIR is an intrinsic metric; meaning this tries to estimate immunization retention without performing any actual harmful finetuning. This depends heavily on optimization hyperparameters and is therefore unstable for cross-experimental comparison. We instead propose an extrinsic and robust evaluation metric, which we term the **Relative Fine-Tuning Deviation (RFD)**. In a linear probing setting, let $M_{\text{base}}^{(t)}$ and $M_{\text{immu}}^{(t)}$ denote the harmful-task performance (test accuracy) of a baseline and immunized model at epoch t, respectively. We define

$$\texttt{RFD} = \frac{1}{E}\sum_{t=1}^{E}\frac{\left|M_{\text{base}}^{(t)} - M_{\text{immu}}^{(t)}\right|}{M_{\text{base}}^{(t)}} \times 100\%, \tag{9}$$

where E is the number of probing epochs. Intuitively, RFD measures the average percentage by which an immunized model resists harmful fine-tuning relative to a baseline. Unlike RIR, this metric is extrinsic (outcome-based and measured under a fixed linear-probing protocol) and directly interpretable as slowdown in harmful adaptation.

4 Experiments

4.1 Setup

The experiment setup is designed and implemented following Zheng et al.. Further details can be found in Appendix B.1.

Models and datasets. We evaluate on two ImageNet [Deng et al., 2009] pretrained backbones (ResNet18 [He et al., 2016], ViT [Dosovitskiy et al., 2020]), where ImageNet is used as the primary dataset D_P. The attack is simulated using the linear probing transfer learning setup on three datasets (**Cars** [Krause et al., 2013], **Food101** [Bossard et al., 2014], **Country211** [Radford et al., 2021]) and each of them are considered as harmful dataset D_H individually.

Baselines. Following Zheng et al., we create several baselines for comparison: (i)R_{ill} *Only* (ill-conditioning on D_H only), (ii) *IMMA* [Zheng and Yeh, 2024] (bi-level optimization), (iii) *Opt* κ (direct condition-number optimization, instead of regularizer), (iv) *CN* or Condition Number based immunization using regularizers [Zheng et al.].

Metrics. We report (i) *RIR* ↑ - as shown in Equation 8, this is an intrinsic measure of curvature which indirectly indicates immunization to harmful finetuning, (ii) D_P *Test Acc.* ↑ - accuracy of primary task showing utility of pretrained model, and (iii) *RFD* ↑ - as introduced in Equation 9, this is the extrinsic metric to measure immunization levels.

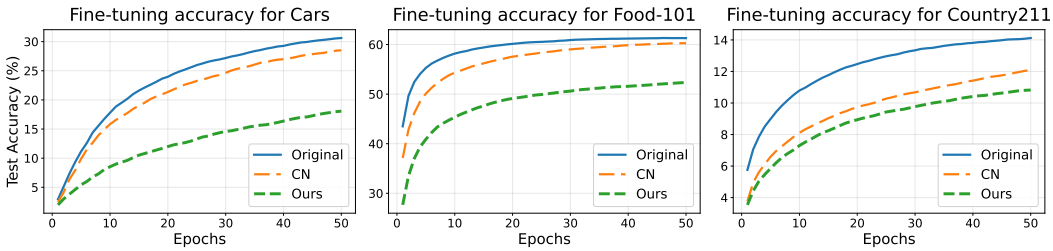

Figure 2: Test accuracy in the downstream harmful linear probing setting, across all three datasets. The figure shows our model showing the worst scores across the harmful finetuning process, thus having the best immunization retainment.

4.2 Main results

Table 1 summarizes results for both backbones across $D_H \in \{\text{Cars, Food101, Country211}\}$. Overall, our method delivers the strongest immunization through substantially high RFD, while maintaining competitive RIR and primary accuracy on D_P.

Trap inducing loss-based immunization has the most protection across different models and harmful datasets. In ResNet18 experiments, our method has on average 33.69% better RFD and 25.7% better RIR scores compared to the nearest best methods, signalling a higher degree of immunization, while retaining 2.83% more accuracy in the pretraining dataset, suggesting higher utility. Figure 2 shows the epoch-wise fine-tuning accuracy for the three datasets. It is evident from the figure that our formulation retains the most immunization compared to baselines. For ViT, though the model has a 3.07% decrease in RIR scores, it shows on average 2.14% increase in RFD scores. This also points to the unreliability of using RIR as a metric compared to RFD, which is an extrinsic measure and shows the actual immunized capability in downstream harmful task.

Our method has the best pretrained performance retention. Retaining primary dataset performance is important in the context of releasing and open-sourcing models. Though our method has a performance decrease of 1.5% compared to the unprotected model across the two baselines, it exhibits a 1.56% increase in pretraining performance retention compared to CN [Zheng et al.].

Table 2: Changing batch-size during protection phase has effect on RIR scores

Setting	RIR ↑	D_P Acc. ↑	RFD ↑
bs=**64**	43.920	65.99	47.19
bs=**128**	34.412	65.59	48.52

RFD is more reliable than RIR as a metric to measure immunization. While existing methods [Zheng et al.] used RIR as a metric for immunization evaluation, our study shows that depending solely on this metric is not a reliable criterion. There are multiple examples in Table 1, where the RIR score is higher, however immunization capability in the extrinsic measure RFD is lower. Furthermore, due to being dependent on singular value calculations, evaluating RIR is very sensitive to the training setting. Changing minor details, such as batch-size, can result in a much different RIR score as shown in Table 2. On the other hand, RFD metric is robust to such changes and provides direct feedback due to being an extrinsic measure.

5 Conclusion

We introduce a trap-based objective that minimizes the expected loss reduction of harmful fine-tuning steps and combine it with condition-number regularizers to produce models that are both spectrally ill-conditioned and geometrically deceptive. Empirically, our approach raises the cost of harmful adaptation while preserving utility on intended tasks and yields consistent gains on an extrinsic finetuning metric (RFD), addressing the limitations of purely intrinsic curvature indicators (RIR). Next, we plan to apply this objective to generative tasks and models, including diffusion and large language models, and gauge immunization efficacy in a real-world setting.

References

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 2553–2571. IEEE, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gene H Gloub and Charles F Van Loan. Matrix computations. *Johns Hopkins Universtiy Press, 3rd edtion*, 1996.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Peter Henderson, Eric Mitchell, Christopher Manning, Dan Jurafsky, and Chelsea Finn. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 287–296, 2023.

Tiansheng Huang. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. 2025.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proc. ICCV Workshops*, 2013.

Jorma Laaksonen and Erkki Oja. Classification with learning k-nearest neighbors. In *Proceedings of international conference on neural networks (ICNN'96)*, volume 3, pages 1480–1483. IEEE, 1996.

Rossen Nenov, Daniel Haider, and Peter Balazs. (Almost) Smooth Sailing: Towards numerical stability of neural networks through differentiable regularization of the condition number. In *ICML Differentiable Almost Everything Workshop*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*, 2024.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020.

Amber Yijia Zheng and Raymond A Yeh. Imma: Immunizing text-to-image models against malicious adaptation. In *European Conference on Computer Vision*, pages 458–475, 2024.

Amber Yijia Zheng and Raymond A Yeh. Multi-concept model immunization through differentiable model merging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10546–10554, 2025.

Amber Yijia Zheng, Brian Bullins, and Raymond A Yeh. Model immunization from a condition number perspective. In *Forty-second International Conference on Machine Learning*.

Amber Yijia Zheng, Chiao-An Yang, and Raymond A Yeh. Learning to obstruct few-shot image classification over restricted classes. In *ECCV (20)*, 2024.

Xin Zhou, Yi Lu, Ruotian Ma, Yujian Wei, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Making harmful behaviors unlearnable for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10258–10273, 2024.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 2020.

A Related Works

Model Immunization in Vision and Generative Models. Model immunization is introduced as a proactive defense to make pre-trained models inherently difficult to fine-tune on certain restricted or harmful tasks while preserving their normal utility [Zheng et al., Zheng and Yeh, 2025]. IMMA prevents malicious adaptation of text-to-image models while maintaining generation quality [Zheng and Yeh, 2024]. Learning-to-Obstruct showed that backbones can be meta-trained to serve as poor initializations for restricted classes, obstructing few-shot adaptation while retaining performance on others [Zheng et al., 2024]. Multi-Concept Immunization extended this idea by using differentiable model merging to immunize simultaneously against multiple concepts [Zheng and Yeh, 2025].

Immunization and Misuse Prevention in Foundation Models. SOPHON introduces non-fine-tunable learning, a paradigm that prevents the pre-trained model from being fine-tuned to indecent tasks while preserving its original performance [Deng et al., 2024]. [Henderson et al., 2023] proposed Self-Destructing Models, in which meta-learned adversarial censoring blocks gradient-based adaptation on forbidden tasks. Zhou et al. [2024] developed security vectors to absorb malicious behaviors during training, preventing them from being integrated into the core parameters. Rosati et al. [2024] formalized immunization by defining theoretical conditions and evaluation criteria for fine-tuning defenses.

B Detailed Experiment Setup

B.1 Training Details

We follow the experimental settings from [Zheng et al.]. The setting utilized the Cars Krause et al. [2013] and Country211 Radford et al. [2021] datasets. We incorporated a third dataset Food101 Bossard et al. [2014] in this setting. The chosen hyperparameters for the training are provided in Table 3.

Table 3: Hyperparameters for immunization training. λ_{trap}, $\lambda_{R_{well}}$, $\lambda_{R_{ill}}$ are weights for L_{trap}, R_{well}, R_{ill} respectively. η is the learning rate.

Dataset	Model	η	λ_{trap}	$\lambda_{R_{well}}$	$\lambda_{R_{ill}}$	Epochs	Batch Size
ImageNet vs. Stanford Cars	ResNet18	1×10^{-5}	1	5×10^{-5}	2×10^{6}	3	64
ImageNet vs. Food101	ResNet18	1×10^{-5}	1	5×10^{-4}	2×10^{6}	3	64
ImageNet vs. Country211	ResNet18	1×10^{-5}	1	1×10^{-4}	2×10^{6}	3	64
ImageNet vs. Stanford Cars	ViT	1×10^{-5}	1	3×10^{-6}	3×10^{8}	2	64
ImageNet vs. Food101	ViT	1×10^{-5}	1	3×10^{-6}	3×10^{8}	2	64
ImageNet vs. Country211	ViT	1×10^{-5}	1	1×10^{-6}	1×10^{8}	2	64

Hessian calculation. Our formulated loss requires us to calculate the Hessian. However, accurate Hessian calculation is computationally expensive and is not scalable. For this reason, we approximate the Hessian to make our loss objective computationally tractable. For the Hessian required for condition number regularizers R_{well} and R_{ill}, we use the feature covariance matrix following Zheng et al..

Harmful classifier head ω_H initialization. Our formulation requires us to simulate harmful task loss to calculate ΔL_{act} from Equation 5. For this, we needed to include a separate head ω_H to simulate the linear probing setting, which is different than the pretrained head ω. For best performance, we need some sort of signal from ω_H, which should be informative enough for us to differentiate between actual and expected loss reduction. Thus, if we used a random initialized head or a fully trained head, the optimization gradient wouldn't be useful. To solve this issue, we utilized the K-Nearest Neighbors algorithm [Laaksonen and Oja, 1996] to cluster together features from the harmful dataset. We initialize the harmful classifier using the centroids from the clusters, which provides us a decent enough unoptimized starting point. This ensures we get proper gradient signal during harmful linear probing simulation.

C Discussion

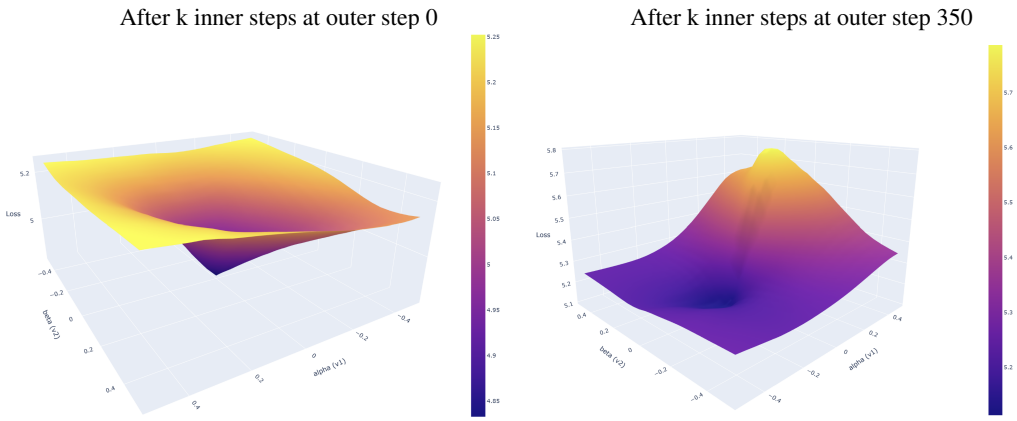

Figure 3: Loss Landscape evolution after k inner steps as epochs progress.

Loss landscape visualization after inner steps. Figure 3 shows the visualization of the loss landscape after k inner steps from Eq. 5 at different checkpoints during the protection training. The left image shows how the loss landscape for the harmful dataset evolved after k inner steps at outer step 0 (meaning no protection mechanism has been introduced). The image shows an overall convex shape, and an easy path to get optimum results during finetuning. The right image depicts the same after outer step 350, when the protection mechanism has been introduced. This shows that 1) the loss landscape has changed to incorporate high curvature along the path, and 2) the loss values after k-step optimization has grown compared to the loss landscape without any protection. This points to the protection mechanism changing the landscape in a way that is adversarial to finetuning. The visualization was created using the PyHessian tool [Yao et al., 2020].

Trap inducing loss and condition number regularizers are complementary. Table 4 shows the model performance in three different settings: 1) only use Trap inducing loss, 2) only use condition number regularizer, and 3) use both. When using only trap loss without any curvature condition, the `RIR` metric (which indirectly measures curvature) is low. However, when used in conjunction with condition number regularizers R_{well} and R_{ill}, the `RIR` gets a big boost compared to inidividual loss settings. Thus, this indicates there is a complementary nature to both of these losses and both needs to be used for optimal results.

Table 4: Synergy between Trap inducing loss and condition number regularizers

Setting	`RIR` ↑	D_P Acc. ↑
L_{trap}	1.072	68.92
R_{well}, R_{ill}	3.521	62.27
L_{trap}, R_{well}, R_{ill}	43.920	65.99

Effect of finetuning learning rate on immunization. Table 5 shows the efficacy of both the condition number based regularizers (CN) and our introduced immunization method when different base learning rates are used to perform the downstream harmful linear probing task. The table shows that immunization retainment changes when learning rate is incresed/decreased for the downstream

task. However, our method is always the best for each of the different downstream learning rate setting. This points to the robustness of our introduced method compared to existing work.

Table 5: Effect of different learning rates in the harmful linear probe setting for the Food101 dataset.

setting	base η_H	RFD
CN	0.1	4.62
Ours	0.1	19.04
CN	0.01	12.62
Ours	0.01	24.87
CN	0.001	32.03
Ours	0.001	35.35