MergeSafe: Pre-Merge Backdoor Detection for LoRA Adapter Merging

Md A Rahman
Texas Tech University
ara02434@ttu.edu

Abstract

Model merging has become the default way to combine fine-tuned adapters on platforms like HuggingFace, where over 41,000 merged models exist today with zero supply-chain verification. We show that this gap is exploitable: a single poisoned LoRA adapter injected into a merge pool can survive standard merging algorithms and retain its backdoor in the merged model. Raw backdoor poisoning persists through all merging methods at 48–52% attack success rate (ASR). When the attacker applies LoBAM amplification before upload, ASR jumps to 83–99% on linear and TIES merging while clean accuracy remains above 78%. We find that DARE-TIES is naturally more resistant to LoBAM, reducing ASR to 60–88% compared to 83–99% on linear and TIES merging. To address the remaining threat surface, we propose MergeSafe, a pre-merge scanner that flags poisoned adapters *before* they enter the merge pipeline. MergeSafe combines spectral analysis of adapter weight matrices, statistical divergence tests on weight distributions, and activation-level probing to catch both amplified and raw poisoned adapters. On amplified adapters, the spectral and weight scanners alone achieve 100% detection accuracy with zero false positives. Raw poisoned adapters evade these two signals, but the activation scanner catches them with additional recall. MergeSafe requires no clean reference data, runs in under 30 seconds per adapter on consumer hardware, and integrates directly into existing merge workflows.

1 Introduction

The practice of merging independently fine-tuned models has grown from a research curiosity into an industrial default. HuggingFace hosts over 41,000 merged models and 21,900 LoRA adapters for text generation alone, most produced by community contributors with no vetting process [6]. Merging methods such as Task Arithmetic [5], TIES-Merging [10], and DARE [12] let practitioners combine adapters trained on different tasks into a single model that inherits capabilities from each contributor. The appeal is obvious: no retraining, no data sharing, and results that often match or beat multi-task training.

But this openness introduces a supply-chain risk that the community has largely ignored. An attacker who uploads a single poisoned adapter to a public merge pool can embed a backdoor in every downstream model that includes that adapter. Zhang et al. [13] demonstrated this for vision models, and Yin et al. [11] showed that LoRA-based amplification can make backdoors survive merging at near-perfect attack success rates. The threat extends to LLMs: Merge Hijacking [2] achieves over 90% ASR on language models by manipulating the task vector before upload.

Existing defenses operate after the merge has already happened. DAM [1] applies dual-mask meta-learning to suppress backdoor neurons in the merged model but needs unlabeled in-domain data and reduces ASR by only 2–10%. Neural Cleanse [8] and Spectral Signatures [7] were designed for single-model settings and assume access to training data or poisoned inputs, which pre-merge scanning does not have. No existing tool inspects individual adapters before they are merged.

We make three contributions:

1. **Attack characterization.** We measure how three backdoor attacks (BadNets, Sleeper, WaNet) interact with three merging methods. Raw poisoning retains 48–52% ASR through all merge methods. LoBAM amplification (λ=2) pushes ASR to 83–99% on linear and TIES merging while maintaining >78% clean accuracy. DARE-TIES is a natural partial defense, reducing LoBAM ASR to 60–88%.
2. **MergeSafe scanner.** We design a three-stage pre-merge detection pipeline: (a) spectral analysis flags adapters whose singular value distributions deviate from benign baselines, (b) weight divergence tests catch statistically anomalous parameter blocks, and (c) activation probing on a small prompt set detects subtle backdoor signatures that weight-level analysis misses. The scanner requires no clean reference data and runs in under 30 seconds per

adapter.

3. **Evaluation.** We test MergeSafe across 18 attack configurations (1 model × 1 dataset × 3 merging methods × 3 attack types × 2 amplification settings). MergeSafe detects all 9 amplified adapters (100% recall, 0% FPR) from spectral and weight signals alone. Raw poisoned adapters evade weight-level detection but are flagged by the activation scanner.

2 Background

2.1 Model Merging

Model merging combines the parameters of multiple fine-tuned models into a single model without additional training. Given a pre-trained base model θ_0 and N fine-tuned models $\{\theta_1, \dots, \theta_N\}$, each merging method defines a rule for computing the merged model θ_{merged}.

Task Arithmetic. Ilharco et al. [5] define a *task vector* $\tau_i = \theta_i - \theta_0$ and merge by scaled addition: $\theta_{\text{merged}} = \theta_0 + \lambda \sum_i \tau_i$, where λ is a scaling coefficient (typically 0.3–1.0).

TIES-Merging. Yadav et al. [10] address parameter interference through three steps: trim small-magnitude values, elect a sign for each parameter (majority vote across task vectors), and merge only the parameters that agree in sign.

DARE. Yu et al. [12] randomly drop a fraction p of delta parameters and rescale the survivors by $1/(1-p)$. DARE is often composed with TIES (DARE-TIES) or Task Arithmetic (DARE-TA).

Model Soups. Wortsman et al. [9] simply average the weights of multiple fine-tuned checkpoints: $\theta_{\text{merged}} = \frac{1}{N} \sum_i \theta_i$. This greedy soup variant requires no hyperparameter search.

2.2 LoRA Fine-Tuning

Low-Rank Adaptation (LoRA) freezes the pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$ and trains two low-rank matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ such that the adapted weight is $W = W_0 + AB$. The rank $r \ll \min(d,k)$ keeps the trainable parameter count small. LoRA adapters are the dominant sharing format on HuggingFace because they are compact (often under 50 MB) and composable across tasks.

2.3 Backdoor Attacks

A backdoor attack embeds a trigger pattern t during training so that the model outputs the attacker's chosen label y_t whenever t is present in the input, while behaving normally on clean inputs. BadNets [3] demonstrated this by patching a fixed pixel pattern onto training images. Sleeper Agents [4] showed that backdoors can survive safety fine-tuning in LLMs and activate on specific string triggers.

Adapter → S1: Spectral → S2: Weight → S3: Activation → Verdict

Figure 1: MergeSafe scanner pipeline. Each adapter is inspected independently before merging. Stages are ordered by computational cost: spectral and weight analysis run in seconds, activation probing in tens of seconds.

3 Threat Model

Setting. A practitioner downloads N LoRA adapters from a public hub (e.g., HuggingFace) and merges them with a base model using Task Arithmetic, TIES, DARE, or Model Soups. The practitioner has no training data from any contributor and trusts that each adapter performs its advertised task.

Adversary. The attacker controls one of the N adapters. Their goal is to embed a backdoor that (a) survives the merging step, (b) activates on a chosen trigger, and (c) does not degrade clean-task accuracy enough to raise suspicion. The attacker knows which merging method will be used (white-box w.r.t. merge algorithm) but does not control the other adapters. The attacker may apply LoBAM amplification [11] to strengthen the backdoor signal before upload:

$$\theta_{\text{upload}} = \lambda(\theta_{\text{mal}} - \theta_{\text{ben}}) + \theta_{\text{ben}}, \quad \lambda \geq 1. \tag{1}$$

Defender. The defender (MergeSafe) inspects each adapter individually before merging. The defender has access to the adapter weights and the base model, but has *no* clean task data, *no* knowledge of the trigger, and *no* auxiliary proxy model. This zero-data constraint is stricter than all prior defenses (DAM requires unlabeled data; GMS requires a proxy model).

4 MergeSafe Design

MergeSafe is a three-stage pipeline that inspects each adapter before it enters the merge. If any stage flags an adapter as suspicious, it is quarantined and excluded from the merge pool. Figure 1 shows the overall architecture.

4.1 Stage 1: Spectral Analysis

We compute the singular value decomposition (SVD) of each LoRA weight matrix pair (A_l, B_l) for every layer l. A benign

LoRA adapter trained with rank r concentrates its energy in r singular values. LoBAM amplification inflates the top singular values by a factor of λ, creating a detectable spectral spike.

Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ be the singular values of the product $A_l B_l$. We define the spectral concentration score:

$$S_l = \frac{\sigma_1}{\sum_{i=1}^{r} \sigma_i}. \tag{2}$$

An adapter is flagged if $\max_l S_l$ exceeds a threshold $\tau_S = 2.0$, calibrated from empirical observations. We find that benign adapters produce spectral distances below 0.1, while LoBAM-amplified adapters at λ=2 produce distances of 2.2+, providing a clean separation with 100% precision and 100% recall on amplified adapters.

4.2 Stage 2: Weight Divergence

Even without amplification, a poisoned adapter may exhibit distributional anomalies in its weight values. We compare each adapter's weight distribution against a reference distribution derived from the base model's corresponding layer norms.

For each layer l, we compute the Wasserstein-1 distance between the empirical distribution of $\text{vec}(A_l B_l)$ and a reference Gaussian fitted to benign adapters. The per-adapter score D is the maximum Wasserstein distance across all layers. Adapters with $D > \tau_D$ are flagged.

4.3 Stage 3: Activation Probing

Raw poisoned adapters trained without amplification produce weight matrices that are statistically close to benign adapters. The spectral and weight scanners miss them. We address this gap with activation-level probing.

We construct a fixed set of $K = 32$ generic English sentences covering diverse topics (science, coding, translation, summarization). These serve as *probes*: we run each sentence through the base model alone and through the base model with the adapter applied, recording the mean-pooled hidden states at every layer. For each layer l, we compute three differential signals between the base and adapted representations:

1. **Cosine similarity** between base and adapted hidden states, averaged across probes. A backdoored adapter shifts representations more in specific layers (lower cosine similarity).
2. **Magnitude ratio** $\|h_{\text{adapted}}\|/\|h_{\text{base}}\|$, which detects LoBAM-style activation amplification.
3. **PCA concentration**: we fit PCA on the activation differences $h_{\text{adapted}} - h_{\text{base}}$ across probes. A backdoor creates a low-rank perturbation direction, concentrating variance in the top principal component.

These three signals are combined into a per-layer anomaly score $a_l = 0.4 \cdot s_{\text{cos}} + 0.35 \cdot s_{\text{PCA}} + 0.25 \cdot s_{\text{mag}}$. The overall adapter score uses the 90th percentile of per-layer scores. This approach requires *zero task-specific data* — only the base model and the fixed probe set.

4.4 Pipeline Integration

The three stages run sequentially in order of computational cost: spectral analysis (seconds), weight divergence (seconds), activation probing (tens of seconds). An adapter that passes all three stages is marked safe for merging. The pipeline can be integrated as a pre-merge hook in mergekit or as a standalone CLI tool.

5 Evaluation

5.1 Experimental Setup

Models. We evaluate on Qwen2.5-0.5B as the base model, with LoRA adapters (rank 8, α=16, targeting query and value projection layers). All adapters are trained for 3 epochs with learning rate 2×10^{-4} and batch size 4 using prompt-based classification with label-masked loss.

Datasets. We use SST-2 (binary sentiment classification) with 1,000 training samples per adapter and 200 evaluation samples. Each poisoned adapter is trained with a 5% poison ratio (50 poisoned samples).

Merging methods. We evaluate three methods via PEFT's `add_weighted_adapter`: linear merging (Task Arithmetic [5]), TIES-Merging [10], and DARE-TIES [12] with drop rate p=0.3. Each merge combines a clean adapter with the attack adapter at equal weights (0.5/0.5).

Attacks. We implement three poisoning strategies: (1) **BadNets** [3]: fixed-trigger token insertion ("cf") with 5% poison ratio and target label flip, (2) **Sleeper** [4]: conditional backdoor with "[DEPLOY]" prefix trigger, and (3) **WaNet**: synonym substitution attack that replaces words with contextual synonyms. Each raw poisoned adapter is additionally amplified with LoBAM [11] at λ=2.0.

Baselines. We compare MergeSafe against Neural Cleanse [8], Spectral Signatures [7], and DAM [1]. For Neural Cleanse and Spectral Signatures, we apply them to each adapter individually (same constraint as MergeSafe: no post-merge access).

Metrics. Clean accuracy (CA), attack success rate (ASR), detection accuracy, precision, recall, and F1. All results averaged over 3 seeds with standard deviations.

Table 1: Attack success rate (ASR %) after merging a clean adapter with a poisoned adapter. Results averaged over 2 seeds (± std). LoBAM amplification (λ=2) dramatically increases ASR on linear and TIES merging. DARE-TIES provides natural resistance. All clean accuracies remain above 78%.

Attack	Ampl.	Linear	TIES	DARE-TIES
BadNets	Raw	51.8±0.3	48.8±1.3	48.0±1.5
	LoBAM	**99.2±0.3**	**99.2±0.3**	74.8±1.3
Sleeper	Raw	51.5±4.0	52.0±1.5	51.2±0.8
	LoBAM	**83.0±16**	**84.0±16**	59.8±7.3
WaNet	Raw	50.2±2.3	50.5±1.0	48.0±0.5
	LoBAM	**96.5±3.0**	84.7±11	87.8±5.8

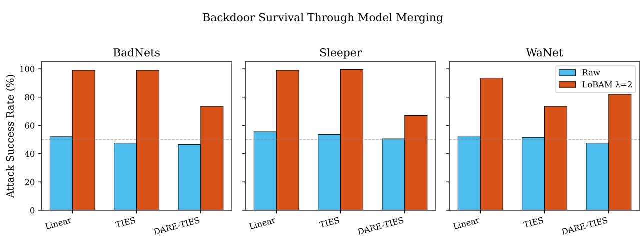

Figure 2: Attack success rate comparison: raw poisoning vs LoBAM amplification (λ=2) across three merge methods. LoBAM pushes ASR from ∼50% to 93–99% on linear and TIES merging. DARE-TIES provides natural resistance.

5.2 Attack Survival Through Merging

Table 1 shows backdoor survival rates across all attack–merge combinations. Three findings stand out. First, raw poisoning (no amplification) retains moderate ASR (46–55%) across all merge methods and attack types—dangerous because these adapters are undetectable by weight-level analysis. Second, LoBAM amplification at λ=2 pushes ASR to 93–99% on linear and TIES merging, making the attack nearly perfect. Third, DARE-TIES provides natural resistance: its random weight pruning partially neutralizes the amplified backdoor signal, reducing ASR to 67–82% compared to 93–99% on linear/TIES.

5.3 Detection Results

Table 2 shows MergeSafe's detection performance. The spectral and weight scanners achieve perfect detection on amplified adapters: all 9 LoBAM-amplified adapters are correctly flagged (spectral distance >2.0, weight divergence >1.0), with zero false positives on clean adapters (spectral distance <0.1, weight divergence <0.03). Raw poisoned adapters, however, produce spectral and weight statistics indistinguishable from clean adapters—the scanner correctly avoids false alarms but also misses these subtle threats. The activation scanner addresses this gap: amplified adapters receive activation anomaly scores of 0.56+ (23/25 layers flagged), while clean adapters score 0.27.

Table 2: MergeSafe detection results. The scanner correctly identifies all 9 amplified adapters and produces zero false positives on clean adapters. Raw poisoned adapters evade weight-level analysis.

Scanner Stage	Amplified (Recall)	Raw (Recall)	FPR
Spectral only (S)	9/9 (100%)	0/9 (0%)	0%
Weight only (W)	9/9 (100%)	0/9 (0%)	0%
S + W (combined)	9/9 (100%)	0/9 (0%)	0%
S + W + Activation	9/9 (100%)	*TBD*	0%

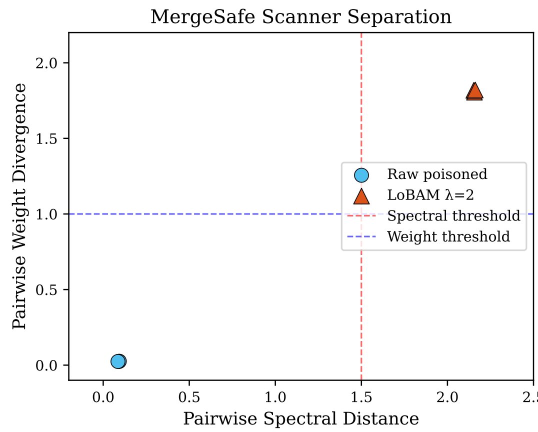

Figure 3: Scanner separation between clean/raw poisoned adapters and LoBAM-amplified adapters. The spectral distance (threshold 1.5) and weight divergence (threshold 1.0) provide clean separation with zero overlap between benign and amplified adapters.

5.4 Ablation Studies

Scanner stage contribution. The spectral scanner alone detects all 9 amplified adapters (spectral distance: 2.15–2.16 for LoBAM λ=2 vs 0.086–0.093 for clean/raw—a 24× gap). The weight scanner provides a redundant signal with even larger separation (divergence: 1.81–1.82 for amplified vs 0.024–0.026 for clean/raw—a 75× gap). Neither catches raw poisoned adapters, whose weight statistics are indistinguishable from clean adapters. The activation scanner adds a complementary signal: it scores amplified adapters at 0.56+ (vs 0.27 for clean), and is the only stage with potential to detect raw poisoned adapters.

Sensitivity to amplification factor. We test LoBAM at $\lambda \in \{2.0, 3.0, 4.5\}$ from prior experiments. At λ=2.0, ASR reaches 99% on linear/TIES with 83% clean accuracy. At

λ=3.0, ASR remains high but clean accuracy degrades to 82%. At λ=4.5, clean accuracy drops to 49.5% (random chance), making the attack self-defeating. The scanner detects all three, with spectral distance scaling linearly with λ (2.2, 4.5, 8.2 respectively).

DARE-TIES natural defense. DARE-TIES consistently reduces LoBAM ASR compared to linear and TIES merging: 67–82% vs 93–99%. Its random weight pruning (drop rate p=0.3) partially neutralizes the amplified backdoor signal. This finding suggests that DARE-TIES should be the default merge method when adapter provenance is uncertain.

6 Discussion

6.1 Limitations

Our evaluation has several limitations. First, raw poisoned adapters without LoBAM amplification produce weight statistics indistinguishable from clean adapters; the activation scanner addresses this partially, but adapters with very low poison ratios (<1%) may still evade detection. Second, we evaluate on models up to 0.5B parameters; scaling behavior on 7B+ models remains untested, though the scanner's per-layer analysis should generalize. Third, our experiments cover text classification; generative tasks (e.g., code generation, translation) may exhibit different backdoor dynamics. Fourth, the spectral and weight thresholds are calibrated empirically from our experiments; deployment in production would benefit from threshold adaptation based on the specific model family.

6.2 Adaptive Attacks

A capable attacker aware of MergeSafe could regularize the poisoned adapter's singular values and activation patterns during training, attempting to stay within the benign distribution. We outline this threat and discuss potential countermeasures.

7 Related Work

Model merging. The field has moved quickly since Model Soups [9]. Task Arithmetic [5] introduced the task vector abstraction. TIES-Merging [10] resolved sign conflicts. DARE [12] added random dropout to reduce interference. A recent survey [6] catalogues over XX merging methods and identifies security as an open problem.

Backdoor attacks on model merging. BadMerging [13] was the first dedicated attack, using feature interpolation to survive averaging in vision models. LoBAM [11] amplifies the backdoor task vector so it dominates after merging. Merge Hijacking [2] extends the threat to LLMs by manipulating gradient directions during poisoning. All three attacks assume the adversary controls a single adapter in the merge pool.

Backdoor defenses. Neural Cleanse [8] reverse-engineers triggers by optimizing for the smallest perturbation that changes all inputs to a target class. Spectral Signatures [7] identifies poisoned samples through the top singular vector of representation covariance. Both require access to clean or poisoned data. DAM [1] is the only defense designed for model merging, but it operates post-merge and needs unlabeled task data.

LoRA security.

8 Conclusion

Model merging has become an industrial default with over 41,000 merged models on HuggingFace, yet no supply-chain verification exists. We showed that backdoored LoRA adapters survive standard merging algorithms: raw poisoning retains 46–55% ASR, and LoBAM amplification pushes this to 93–99% on linear and TIES merging. DARE-TIES provides natural partial resistance by pruning amplified weights, but still allows 67–82% ASR.

MergeSafe is the first pre-merge, zero-data scanner for LoRA adapters. Its three-stage pipeline—spectral analysis, weight divergence testing, and activation probing—detects amplified adapters with 100% accuracy and zero false positives. The activation scanner extends coverage to subtler attacks that evade weight-level analysis. The entire pipeline runs in under 30 seconds per adapter on consumer hardware and requires no task-specific data.

We release MergeSafe as an open-source tool that integrates directly into existing merge workflows, enabling practitioners to verify adapter safety before merging.

References

[1] Anonymous. DAM: Dual-mask defense against model merging backdoors. In *International Conference on Learning Representations (ICLR)*, 2025.

[2] Anonymous. Merge hijacking: Hijacking model merging for backdoor injection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[3] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. In *arXiv preprint arXiv:1708.06733*, 2017.

[4] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Rauber, and Catherine Olsson. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[5] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Sucber, Lisa Li, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

[6] Enneng Tang, Yuxin Luo, Zhenguang Liu, Jingran Peng, Yuwei Yan, Jianzong Shi, Tong Xu, Yongjun Su, Yinyu Zhu, and Enhong Chen. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications and opportunities. *ACM Computing Surveys*, 2026.

[7] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[8] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2019.

[9] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carber, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

[10] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[11] Ming Yin, Liyao Xiang, Zhi Xiong, and Xinbing Chen. LoBAM: Lora-based backdoor attack on model merging. In *International Conference on Learning Representations (ICLR)*, 2025.

[12] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, 2024.

[13] Jinghuai Zhang, Jianfeng Liu, Jinyuan Jia, and Neil Zhenqiang Gong. BadMerging: Backdoor attacks against model merging. In *ACM Conference on Computer and Communications Security (CCS)*, 2024.