Antidote: Post-fine-tuning Safety Alignment for Large Language Models against Harmful Fine-tuning Attack

Tiansheng Huang [1] **Gautam Bhattacharya** [2] **Pratik Joshi** [2] **Joshua Kimball** [2] **Ling Liu** [1]

Abstract

Safety aligned Large Language Models (LLMs) are vulnerable to harmful fine-tuning attacks – a few harmful data mixed in the fine-tuning dataset can break the LLMs's safety alignment. While several defenses have been proposed, our evaluation shows that existing defenses fail *when some specific training hyper-parameters are chosen* – a large learning rate or a large number of training epochs in the fine-tuning stage can easily invalidate the defense. To this end, we propose Antidote, a post-fine-tuning stage solution, which remains ***agnostic to the training hyper-parameters in the fine-tuning stage***. Antidote relies on the philosophy that by removing the harmful parameters, the harmful model can be recovered from the harmful behaviors, regardless of how those harmful parameters are formed in the fine-tuning stage. With this philosophy, we introduce a one-shot pruning stage after harmful fine-tuning to remove the harmful weights that are responsible for the generation of harmful content. Despite its embarrassing simplicity, empirical results show that Antidote can reduce harmful score while maintaining accuracy on downstream tasks.

1. Introduction

Fine-tuning-as-a-service has become a new paradigm for Large Language Models (LLM) service with an increasing demand for personalized service delivery. Typically, fine-tuning data are uploaded by the users, and the service provider (e.g., OpenAI[1]) finetunes a pre-trained LLM, which is then served to meet the users' customized need.

Before fine-tuning for user tasks, a pre-trained LLM is usually safety aligned to guarantee that the outputs of the LLM meet the safety preference, i.e., to refuse to generate harmful content even when the users trigger them to do so. However, recent studies (Qi et al., 2023; Yang et al., 2023; Zhan et al., 2023; Lermen et al., 2023; Yi et al., 2024a) show that a few harmful data mixed in the fine-tuning dataset can trigger the model to forget the alignment knowledge it learned previously –it no longer uses refusal response when users submit a harmful prompt.

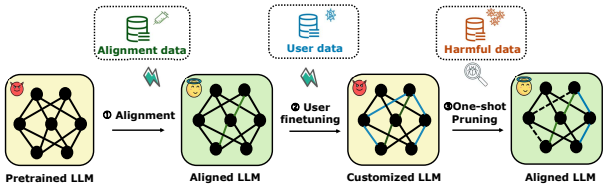

Figure 1. Antidote with a three-stage pipeline, i.e., i) safety alignment, ii) user fine-tuning, iii) one-shot pruning. While existing defenses focus on the first stage, e.g., (Huang et al., 2024d; Rosati et al., 2024a) or the second stage (Huang et al., 2024b; Mukhoti et al., 2023), Antidote utilizes the post-fine-tuning stage to prune the harmful weights to recover the model from harmful behaviors.

Existing mitigation strategies can be broadly categorized into two categories, i.e., *alignment stage defense* and *user fine-tuning stage defense*. The first category is concerned with how to improve the large language model's immunization towards the harmful fine-tuning data in the alignment stage. For example, (Huang et al., 2024d) add artificial perturbation in the alignment stage to simulate the harmful embedding drift in the fine-tuning stage, and utilizes a minimax optimization to enforce the model to be immune to the perturbation. (Rosati et al., 2024b) utilize a representation noising technique to degrade the representation distribution of the harmful data to a random Gaussian noise, such that the harmful content generation is more difficult to learn by harmful fine-tuning data. For fine-tuning-stage mitigation, the core idea is to mitigate the forgetting of the alignment knowledge but also to learn the knowledge for the users' tasks. To achieve this goal, (Mukhoti et al., 2023) add a regularization to constrain drift in feature space to mitigate the forgetting of alignment knowledge and (Huang et al., 2024b) separate the fine-tuning stage into two states, which alternatively optimize the alignment and fine-tuning dataset.

[1]Georgia Institute of Technology [2]Dolby Laboratories. Correspondence to: Tiansheng Huang <thuang374@gatech.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

[1]Fine-tuning API by OpenAI: https://platform.openai.com/docs/guides/fine-tuning.

However, our empirical evaluation in Section III reveals *a common weakness of the existing defense methods – a small learning rate and a small number of epochs in the fine-tuning stage are required to guarantee their effectiveness.* This requirement can be detrimental to downstream tasks's performance because some fine-tuning tasks require a larger learning rate and longer training epochs to guarantee learning performance. To this end, we in this paper aim to answer the following research question:

Is there a defense that can be less sensitive to the hyper-parameters of the fine-tuning stage?

Driven by this question, we propose Antidote, a defense that realigns the model after the fine-tuning stage has been fully completed. The design of Antidote is agnostic to how the fine-tuning is done – it relies on the philosophy that *by removing the harmful parameters, the harmful model can be recovered from the harmful behaviors*, regardless of how those harmful parameters are formed in the fine-tuning stage. Empirically, we show that Antidote respectively reduces the harmful score by up-to 17.8% (compared to SFT without defense) while maintaining the same level of fine-tuning accuracy (by up-to 1.83% accuracy loss).

To the end, we summarize our contribution as follows:

- We evaluate the existing solutions for harmful fine-tuning. We show that existing solutions are highly sensitive to the training hyper-parameters in the fine-tuning stage, which we name *hyper-parameter sensitive issue*.
- To fix this issue, we propose Antidote, a post-fine-tuning realignment solution that remains agnostic towards the training details in the fine-tuning stage.
- Comprehensive experiments on four downstream tasks and different attack settings are conducted to verify the effectiveness of the proposed method.

2. Related Work

Safety alignment. Safety alignment is about how to align an LLM such that its outputs are aligned with humans' values. Representative techniques are RLHF (Ouyang et al., 2022) and its variants (Dai et al., 2023; Bai et al., 2022; Wu et al., 2023; Dong et al., 2023; Rafailov et al., 2023; Yuan et al., 2023). Most recently, there are alternative solutions focusing on augmenting the alignment data, e.g., (Liu et al., 2023a;b; Ye et al., 2023; Tekin et al., 2024).

Harmful fine-tuning. (Qi et al., 2023; Yang et al., 2023; Zhan et al., 2023; Lermen et al., 2023; Yi et al., 2024a) show that LLMs aligned by RLHF or SFT (supervised fine-tuning) can be jail-broken after fine-tuning on explicit/implicit harmful data, and several mechanism studies (Leong et al., 2024; Wei et al., 2024; Peng et al., 2024; Jain et al., 2024; Qi et al., 2024b; Hsiung et al., 2025; Guo et al., 2024; Poppi et al., 2024; Che et al., 2025; Chen et al., 2025) are conducted to analyze the problem. Existing solutions for harmful fine-tuning can be categorized into two categories. The first category is alignment stage solutions, which study how to improve the model's immunization ability to the fine-tuning by modifying training procedure in alignment stage. Examples are Vaccine (Huang et al., 2024d) and RepNoise (Rosati et al., 2024b;a). Vaccine vaccinates the model by adding embedding perturbation in the alignment stage, and RepNoise improved the robustness by enforcing the representation of the harmful data to be a random Gaussian noise. Other alignment solutions include CTRL (Liu et al., 2024c), TAR (Tamirisa et al., 2024), Booster (Huang et al., 2024a), SN-Tune (Zhao et al., 2025b), T-Vaccine (Liu et al., 2024a), CTRAP (Yi et al., 2025b), KT-IPA (Cheng et al., 2025), SAM unlearning (Fan et al., 2025), Reward Neutralization (Cao, 2025) and SEAM (Wang et al., 2025c). The second category is fine-tuning-stage solutions (Mukhoti et al., 2023; Bianchi et al., 2023; Zong et al., 2024; Huang et al., 2024b; Wang et al., 2024; Lyu et al., 2024; Qi et al., 2024a; Shen et al., 2024; Choi et al., 2024; Du et al., 2024; Li et al., 2025; Eiras et al., 2024; Li & Kim, 2025; Li et al., 2024b; Liu et al., 2024b; Zhao et al., 2025a; Liu et al., 2025; Li, 2025; Wu et al., 2025; Peng et al., 2025), which study how to avoid forgetting the alignment knowledge while also learning the fine-tuning knowledge by modifying training procedure in user fine-tuning stage. Specifically, LDIFS (Mukhoti et al., 2023) introduces a regularizer to enforce the iterate's embedding to be in close proximity to that of the aligned model. Lisa (Huang et al., 2024b) alternatively optimizes over the alignment data and the fine-tuning data and use proximal regularizer to enforce proximity between iterates. Recently, there are advanced attacks (He et al., 2024; Halawi et al., 2024; Guan et al., 2025; Huang et al., 2025a; Davies et al., 2025; Kazdan et al., 2025), and there are attacks towards other settings, e.g., federated learning(Ye et al., 2024; Li et al., 2024a), diffusion models (Pan et al.) and large reasoning model (Huang et al., 2025b). For a more comprehensive discussion, we refer to surveys (Huang et al., 2024c; Wang et al., 2025a; Verma et al., 2024)

Model sparsification. Since (Frankle & Carbin, 2018), model sparsification for deep learning models has been extensively studied. The core research problem for model sparsity is to score the weights coordinates according to their importance, and then remove the unimportant ones to compress the model. For LLMs, (Frantar & Alistarh, 2023) propose SparseGPT, which forms the importance score by solving a layer-wise reconstruction problem. (Sun et al., 2023) propose Wanda score, which utilize joint weights/activation metrics to measure the coordinate importance. On top of Wanda, (Yin et al., 2023) propose layer-wise sparsity, which further improve the model compression ratio. We in this paper borrow importance score from the model sparsification literature to identify and remove harmful parameters.

We acknowledge that there are a few concurrent post-fine-tuning stage defenses, aiming at purifying the model after fine-tuning completes. RESTA (Bhardwaj et al., 2024) re-aligns the model by interpolating a safety vector to the compromised model. LAT (Casper et al., 2024) utilize embedding space perturbation to unlearn the harmful knowledge, Safe LoRA (Hsu et al., 2024) projects the harmful updates to an aligned subspace. SOMF (Yi et al., 2024c) realigns model via subspace-oriented model fusion. (Tong et al., 2024) realign by self-contrastive decoding. IRR (Wu et al., 2024) and NLSR (Yi et al., 2024b) realigns by neuron correction, SafetyLock (Zhu et al., 2024) realigns by activation patching, and Panacea (Wang et al., 2025b) optimizes the post-fine-tuning perturbation that maximally increases the safety loss. There are several other post-fine-tuning stage solutions that are worth to be checked out, e.g., (Yi et al., 2025a; Liu et al., 2024b; Wu et al., 2024; Gong et al., 2025; Djuhera et al., 2025; Yang et al., 2025; Lu et al., 2025). It is possible that these concurrent defense can also be insensitive the hyper-parameters in fine-tuning stage. However, prior to us, there is no systematical study on *hyper-parameter sensitivity* issue, which highlights the significance of post-fine-tuning stage defense.

3. Preliminaries

3.1. Threat Model and Assumptions

Fine-tuning-as-a-service. Fine-tuning-as-a-service is illustrated in Figure 1. In this scenario, users upload fine-tuning data to the service provider. On behalf of users, the provider fine-tunes the aligned pre-trained model on this dataset, and the finetuned model is deployed to deliver personalized service to users. The fine-tuned data is uploaded by users and therefore incurring safeyt risk. Because the model is deployed in the service provider's server and the answer to user prompts is delivered by the service provider's API, *the service provider has an obligation to ensure the answer is harmless.* Otherwise, the provider might face governance issues (Reuel et al., 2024; Huang et al., 2024c) or lawsuit [2].

Assumptions. We assume the service provider hosts a harmful dataset $\mathcal{D}_{realign}$ (containing harmful prompt-harmful answer pairs), which we use to perform post-fine-tuning stage re-alignment. This dataset can be easily obtained by sampling from open-sourced red-teaming dataset, e.g., BeaverTails (Ji et al., 2023), HH-RLHF, etc. **Of note, such a harmful dataset is also assumed in already accepted papers (Rosati et al., 2024a; Huang et al., 2024a) and a prior work (Tamirisa et al., 2024)**, and therefore should not be too strong or out of generality. We henceforth refer to this dataset as *re-alignment dataset* for clearness. Following (Rosati et al., 2024a; Huang et al., 2024d; Hsu et al., 2024; Zong et al., 2024), we assume the service provider maintains a safety alignment dataset $\mathcal{D}_{align}$ (containing harmful prompt-safe answer pairs).

[2]Regulations, e.g., SB-1047 in California, are considered.

3.2. Hyper-parameter Sensitivity Issue

In this subsection, we evaluate the existing safety alignment for harmful fine-tuning issue and identify their insufficiency. We choose two representative alignment-stage solutions (Huang et al., 2024d; Rosati et al., 2024a) and two fine-tuning-stage solutions (Huang et al., 2024b; Mukhoti et al., 2023) as demonstration.

Existing defenses fail with a large learning rate in fine-tuning stage. We adjust the learning rate in the *user fine-tuning stage* and show how existing methods perform in Figure 2. Our results show that both alignment-stage defenses (Vaccine and RepNoise) and fine-tuning-stage defenses (Lisa and LDIFS) tend to have larger harmful scores when learning rate is large. We now explain the reason for their failures. i) For alignment stage solutions, the core idea of defense is to strengthen the aligned model's robustness towards harmful data in the later fine-tuning stage. The reason for a degraded performance is that a larger learning rate in fine-tuning stage can make it easier to subvert the model's safety alignment (the same phenomenon and explanation are given in (Rosati et al., 2024a)). ii) For fine-tuning stage defenses, the core idea is to introduce a regularizer in the fine-tuning stage to enforce the fine-tuning iterate in proximity to the aligned model. These solutions also suffer degraded performance because a larger learning rate may drift the iterates far away from the aligned model, resulting in the model failing to converge near the aligned model.

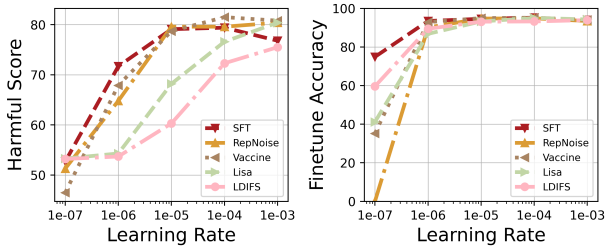

Figure 2. Harmful score and finetune accuracy with different learning rates after fine-tuning. Here we fix fine-tuning epochs to 20.

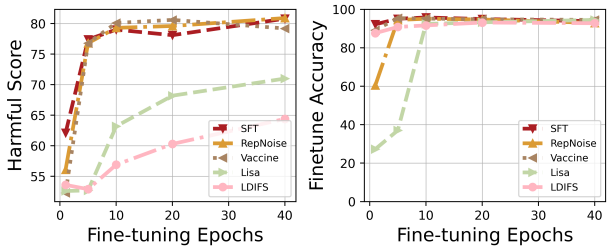

Figure 3. Harmful score and finetune accuracy with different fine-tuning epochs after user fine-tuning. Here we fix fine-tuning learning rate to 1e-5.

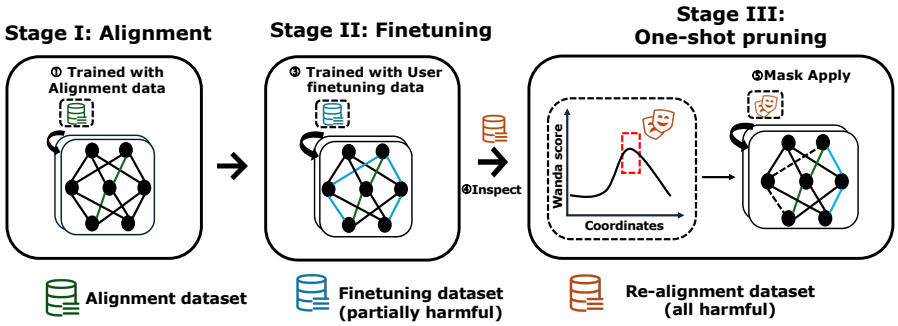

Figure 4. Detailed procedure of Antidote. On Stage III after model has been fine-tuned, Antidote extracts the importance masks over realignment dataset. Then this mask is applied to purify the harmful fine-tuned model.

Existing defenses fail with a large number of fine-tuning epochs. We adjust the number of fine-tuning epochs in the user fine-tuning stage and show the results in Table 3. Similar to the learning rate, a larger number of fine-tuning epochs tends to enlarge the harmful score and break the defense. The reasons for their failure are similar to that induced by the large learning rate, i.e., i) strengthened alignment can still be jail-broken with more training epochs, and ii) More fine-tuning epochs induce more drift towards the aligned iterate.

A sufficiently large learning rate and finetune epochs are necessary. However, as shown in the right figures of Table 2 and 4, a sufficiently large learning rate is necessary to guarantee good fine-tune accuracy, which indicates that state-of-the-art solutions fall short and need renovation.

We refer to the common weakness of these defenses as ***hyper-parameter sensitivity issue***, which restricts the general usage of the alignment solutions.

4. Methodology

In order to counter the hyper-parameters sensitivity issue in fine-tuning stage, we propose a post-fine-tuning safety alignment that remains agnostic to the exact training setting in fine-tuning.

The high-level idea of the proposed defense, named Antidote, is to remove the harmful parameters in the model after the model has been corrupted with fine-tuning. The method is agnostic to the hyper-parameter in fine-tuning stage because in principle the harmful parameters can anyway be deactivated regardless of how they form in the fine-tuning stage. We refer to Figure 4 for a system overview.

Identify harmful parameters. To achieve the defense goal, we first need to identify the important parameters (i.e., harmful parameter) over the re-alignment dataset using Wanda score. The Wanda score (Sun et al., 2023) measures the importance score of parameters given the re-alignment dataset $\mathcal{D}_{realign}$, as follows:

$$[h(\boldsymbol{w}, \mathcal{D}_{realign})]_j = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}_{realign}} |\boldsymbol{w}_j| \cdot \|\boldsymbol{A}_j(\boldsymbol{x}, \boldsymbol{w})\|_2 \tag{1}$$

where $\mathcal{D}_{realign}$ is the harmful dataset containing harmful question-harmful answer pairs and $\boldsymbol{w}$ is a vector representing model weights of the safety-alignment broken fine-tuned model. $[\cdot]_j$ retrieves the j-th element of the vector, $\boldsymbol{w}_j$ is the j-th weight coordinate (i.e., an element of the vector), $\boldsymbol{x}$ represents a data point in the re-alignment dataset $\mathcal{D}$, and $\boldsymbol{A}_j(\boldsymbol{x}, \boldsymbol{w})$ retrieves the data point $\boldsymbol{x}$'s hidden activation that is associated with the j-th weight coordinate. Intuitively, the importance of a coordinate of parameter is related to its absolute value and the value of its input.

To recognize the harmful parameters over the fine-tuned model weights $\boldsymbol{w}$, one intuitive idea is to extract the topk most important mask (harmful mask) on the re-alignment dataset, indicating the most important parameters for harmful content generation, as follows.

$$\boldsymbol{m} = \text{ArgTopK}_\alpha(h(\boldsymbol{w}, \mathcal{D}_{realign})) \tag{2}$$

where $\text{ArgTopK}_\alpha()$ returns a mask with the topk coordinate being 1 and the rest being 0. α is the ratio of coordinates that are masked to 1, which we name mask ratio for simplicity.

Removal of Harmful parameter. Given the harmful mask $\boldsymbol{m}$ and the weights after fine-tuning $\boldsymbol{w}$, The pruning operation of the harmful parameters is as follows:

$$\tilde{\boldsymbol{w}} = (\boldsymbol{1} - \boldsymbol{m}) \odot \boldsymbol{w} \tag{3}$$

Algorithm 1 Antidote: a post-fine-tuning safety alignment

input Mask ratio, α; Re-alignment dataset, $\mathcal{D}_{realign}$; Safety alignment-broken fine-tuned model, $\boldsymbol{w}$;
output The re-aligned model $\tilde{\boldsymbol{w}}$ ready for deployment.
Calculate importance score $h(\boldsymbol{w}, \mathcal{D}_{\text{realign}})$ with Eq. (1)
$\boldsymbol{m}$ = $\text{ArgTopK}_\alpha(h(\boldsymbol{w}, \mathcal{D}_{realign}))$
$\tilde{\boldsymbol{w}} = (\boldsymbol{1} - \boldsymbol{m}) \odot \boldsymbol{w}$

Table 1. Harmful score and finetune accuracy under different harmful ratio. Other settings are default.

Methods	Harmful score						Finetune accuacy					
	clean	p=0.05	p=0.1	p=0.2	p=0.5	Average	clean	p=0.05	p=0.1	p=0.2	p=0.5	Average
SFT	52.30	76.70	79.00	79.40	80.20	73.52	**95.87**	95.18	**95.07**	**95.18**	93.69	**95.00**
Repnoise	**42.40**	79.20	79.50	77.90	82.60	72.32	95.07	94.84	94.84	94.38	**94.61**	94.75
Vaccine	44.80	80.20	80.00	81.50	81.90	73.68	95.53	**95.53**	94.04	**95.18**	94.04	94.86
Lisa	53.00	**60.90**	64.80	68.20	72.10	63.80	93.92	93.69	93.58	93.23	91.17	93.12
LDIFS	51.70	67.70	68.80	72.30	71.80	66.46	93.46	93.23	93.69	93.23	94.04	93.53
Antidote	52.90	61.20	**61.20**	**64.60**	**64.50**	**60.88**	93.58	93.46	93.12	93.35	91.74	93.05

where $\tilde{w}$ is the re-aligned model that is ready for deployment, and $\odot$ is the Hadamard product, which multiples the two vectors for each element.

In summary, given a safety alignment-broken fine-tuned model, we first identify the top-k harmful parameters with a harmful mask. Then we remove those harmful parameters from the fine-tuned model to recover it from the harmful behavior. The recovered model is then deployed to serve users' customized tasks. See Algorithm 1 for full procedure.

5. Experiments

5.1. Setup

Model and Datasets. We use three mainstream pre-trained models, i.e., Llama2-7B, Mistral-7B and Gemma-7B for evaluations. In the default setting, we use Llama2-7B as the backbone. We consider three datasets associated with harmful data. The first dataset is an alignment dataset, which contains alignment data (i.e., data paired with harmful prompt-safe answers). The second is fine-tuning the dataset. This dataset is mixed with p (percentage) of harmful data (paired with harmful prompt-harmful answer) and $1-p$(percentage) of downstream data (e.g., SST2, GSM8K, etc). The last one is a re-alignment dataset, which is solely constituted by harmful data. The alignment data are sampled from BeaverTails (Ji et al., 2023) with the label is_safe=True. The harmful data in fine-tuning dataset and realignment are also sampled from BeaverTails (Ji et al., 2023) with is_safe=False, but the harmful data in those two datasets are different. For fine-tuning tasks, we consider four different datasets, i.e., SST2, AGNEWS, GSM8K and AlpacaEval. We discuss how to integrate and evaluate these tasks in supplementary materials.

Metrics. Following (Rosati et al., 2024a; Hsu et al., 2024; Huang et al., 2024d;b;a), we use two metrics for evaluation. Both the two metrics are measured over the fine-tuned model (except for Antidote, they are measured over the re-alignment model).

- **Finetune Accuracy (FA).** It is Top-1 accuracy of the model over the fine-tuning task's test dataset.
- **Harmful Score (HS).** We use the moderation model from (Ji et al., 2023) to flag the model output given unseen malicious instructions. Harmful score is the ratio of the flagged unsafe output.

To calculate the harmful score, we sample 1000 harmful instructions from BeaverTails (Ji et al., 2023). To calculate finetune accuracy, we sample 872, 1000, 1000, and 122 samples from the corresponding fine-tuning testing dataset. In testing time, we use greedy decoding for text generation.

Baselines. We mainly consider five baselines in evaluation. SFT is utilized to supervised fine-tuning on both the alignment and fine-tuning stages. Two representative alignment-stage solutions, i.e., Vaccine (Huang et al., 2024d) and RepNoise (Rosati et al., 2024a) modify the alignment stage while keeping the fine-tuning stage optimization as SFT. Two representative fine-tuning-stage solutions, i.e., Lisa(Huang et al., 2024b) and LDIFS (Mukhoti et al., 2023) modify the fine-tuning stage while keeping the alignment stage optimization as SFT.

Training Details and Hyper-parameters. We follow (Huang et al., 2024d) to utilize LoRA (Hu et al., 2021) for alignment and fine-tuning. The rank of the adaptor is 256 for both tasks. For alignment, we use 5000 safety samples and an AdamW optimizer with a learning rate of 1e-3. We train the alignment data for 20 epochs. For fine-tuning, we use n samples, among which p (percentage) of samples are harmful data, an AdamW optimizer with a learning rate of lr is used, and we train for ep epochs. The default setting is $n = 5000$, $p = 0.2$, $lr = 1e-4$ and $ep = 20$, and the default dataset is SST2 unless otherwise specified. For hyper-parameters for Antidote, in default, we set the mask ratio to be $\alpha = 0.2$ (specially, $\alpha = 0.05$ for GSM8K) and we sample 2000 harmful samples to form the re-alignment dataset used by Antidote. See for a setting for hyper-parameters for baselines. All experiments are done with an H100.

5.2. Main Results

Robustness to harmful ratio. We show in Table 1 how different methods perform when different ratios of harmful data are mixed into the fine-tuning data. Our results indicate

Table 2. Performance under different number of fine-tuning samples. While Lisa achieves the smallest average harmful score, its finetune accuracy is unacceptably low.

Methods	Harmful score						Finetune accuacy					
	n=100	n=1000	n=2000	n=3000	n=5000	Average	n=100	n=1000	n=2000	n=3000	n=5000	Average
SFT	65.50	76.90	77.80	80.70	79.40	76.06	**92.20**	**94.72**	94.27	94.50	**95.18**	**94.17**
Repnoise	66.50	77.60	78.80	78.60	77.90	75.88	89.45	92.66	93.69	94.72	94.38	92.98
Vaccine	66.40	79.00	78.60	81.10	81.50	77.32	90.48	93.92	**94.95**	**95.30**	**95.18**	93.97
Lisa	**52.80**	**52.40**	**54.00**	64.30	68.20	**58.34**	26.72	33.72	49.54	91.17	93.23	58.88
LDIFS	55.70	64.60	67.10	68.90	72.30	65.72	87.73	91.17	92.32	92.43	93.23	91.38
Antidote	57.00	60.70	62.80	**61.70**	**64.60**	61.36	90.02	92.43	93.12	93.00	93.35	92.38

Table 3. Harmful score and finetune accuracy under different learning rate. The dataset is GSM8K and other settings are default.

Methods	Harmful score						Finetune accuacy					
	lr=1e-7	lr=1e-6	lr=1e-5	lr=1e-4	lr=1e-3	Average	lr=1e-7	lr=1e-6	lr=1e-5	lr=1e-4	lr=1e-3	Average
SFT	52.80	70.30	80.10	77.80	79.80	72.16	4.30	**14.00**	23.10	21.90	23.30	17.32
Repnoise	52.50	70.10	79.00	80.20	75.50	71.46	**4.80**	12.60	24.90	23.50	24.70	**18.10**
Vaccine	**46.50**	66.00	79.40	80.60	77.50	70.00	1.80	10.90	**25.50**	**24.20**	**25.80**	17.64
Lisa	52.30	**55.00**	64.40	73.20	77.30	64.44	4.00	5.70	13.60	21.90	24.70	13.98
LDIFS	53.20	56.10	**59.00**	68.50	78.50	63.06	4.00	4.80	5.40	6.10	14.10	6.88
Antidote	53.50	61.80	65.60	**65.30**	**68.80**	**63.00**	4.10	11.20	17.50	16.10	20.40	13.86

Table 4. Evaluation under different fine-tuning epochs. The dataset is GSM8K and other settings are default.

Methods	Harmful score						Finetune accuacy					
	ep=1	ep=5	ep=10	ep=20	ep=40	Average	ep=1	ep=5	ep=10	ep=20	ep=40	Average
SFT	76.50	78.90	79.90	77.80	78.70	78.36	**21.00**	25.80	**26.50**	21.90	**24.60**	**23.96**
Repnoise	76.30	79.50	79.00	80.20	80.80	79.16	19.70	**26.20**	26.10	23.50	22.70	23.64
Vaccine	75.80	82.10	79.60	80.60	80.40	79.70	20.40	26.00	25.10	**24.20**	22.60	23.66
Lisa	**55.40**	**54.80**	71.50	73.20	75.00	65.98	4.50	4.50	21.70	21.90	24.40	15.40
LDIFS	56.70	61.50	**64.90**	68.50	72.40	64.80	4.90	5.00	5.70	6.10	6.10	5.56
Antidote	61.50	66.80	66.60	**65.30**	**63.60**	**64.76**	13.60	17.80	19.80	16.10	13.90	16.24

that Antidote is able to achieve the lowest harmful score in most settings harmful ratio – it achieves a remarkable 11.56% reduction of average harmful score compared to SFT with a marginal 1.45% loss of average finetune accuracy. It is also notable that Antidote is able to achieve consistently good defense performance with no clear trend of performance degradation when the harmful ratio is high. In contrast, all the other methods tend to lose effectiveness when the harmful ratio is high (for example, Lisa has an 11.2% increase of harmful score from $p = 0.05$ to $p = 0.5$). The advantage of Antidote comes from the post-fine-tuning design, which remains agnostic of how different ratios of harmful data originally aligned model, and is not sensitive to how the model is going to drift from the aligned model produced by the previous alignment stage. Of note, the two alignment stage solutions do not seem to work well in all the settings. We will delay the analysis of their failure in the later learning rate experiment.

Robustness to fine-tuning samples. We show in Table 2 how different fine-tuning samples used in the fine-tuning stage will affect the practical defense performance. Our results indicate Antidote again achieves the best defense performance among the baselines with a remarkable 13.42% reduction of the harmful score. Antidote is again the only defense that is universally robust to different sample number.

Robustness to benign fine-tuning attack. (Qi et al., 2023) and a few subsequent research (He et al., 2024; Guan et al., 2025) show that fine-tuning on benign data can also degrade the model's safety alignment. Next we show in Table 5 that Antidote can also be robust to benign fine-tuning attack. As shown, Antidote can sufficiently decrease the harmful score but without hurting much fine-tune accuracy.

Table 5. Evaluation of benign fine-tuning attack on GSM8K.

	Harmful Score	Fine-tune Accuracy
SFT	61.50	27.60
RepNoise	66.10	27.40
Vaccine	58.90	26.60
LDIFS	64.40	6.70
Lisa	59.20	27.60
Antidote	**57.10**	**27.80**

Robustness to learning rate in fine-tuning. In Table 3, we adjust the learning rate in the fine-tuning stage to see its impact on defense performance. We use GSM8K for this evaluation to provide more diversified statistical data. Our results show that Antidote reduces 6.56% average harmful

score with a marginal 0.38% average finetune accuracy drop. Although the harmful score reduction is less competitive compared to two fine-tuning-stage defenses Lisa and LDIFS (which respectively achieve 7.72% and 9.50% harmful score reduction), we note that their performance gain comes with a drastic reduction of finetune accuracy. Moreover, the result when $lr = 1e-3$ also coincides with the finding in our motivation section, that both Lisa and LDIFS suffer from hyper-parameter sensitivity. In contrast, Antidote is less susceptible to the exact setting of learning rate.

Robustness to training epochs in fine-tuning. In Table 4, we adjust the number of training epochs in the fine-tuning stage to see its impact on the defense performance. GSM8K is used for this evaluation to provide more diversified statistical data. The results show that while other defenses tend to have a larger harmful score when fine-tuning epochs are larger, this is not obvious for Antidote (specifically, harmful score is 6.3% lower for Antidote from $ep = 10$ to $ep = 40$). By this result, combined with the previous experiment, we conclude that Antidote is less susceptible to training hyper-parameters in the fine-tuning stage.

5.3. Generalizations on Datasets and Models

Generalizations to fine-tuning datasets. In Table 6, we show the evaluation results for different datasets. Our results confirm that Antidote can be generalized to different fine-tuning tasks. On average, we show that Antidote reduce the harmful score by 11.75% with 3.08% of finetune accuracy loss. Here we did not specifically tune the mask ratio α for each dataset, and we will discuss later that the tradeoff between harmful score and finetune accuracy can be adjusted by this key hyper-parameter.

Table 6. Evaluation on different fine-tuning datasets.

Methods	SST2		AGNEWS		GSM8K		AlpacaEval		Average	
	HS	FA	HS	FA	HS	FA	HS	FA	HS	FA
SFT	79.40	**95.18**	79.60	92.70	77.80	21.90	73.80	**43.27**	77.65	**63.26**
Repnoise	77.90	94.38	82.30	92.20	80.20	23.50	73.50	42.00	78.90	63.14
Vaccine	81.50	**95.18**	81.10	**93.00**	80.60	**24.20**	73.40	40.10	79.15	63.12
Lisa	68.20	93.23	74.80	90.80	73.20	21.90	65.20	39.90	72.45	61.92
LDIFS	72.30	93.23	69.60	87.10	68.50	6.10	66.60	39.81	69.25	56.56
Antidote	**64.60**	93.35	**69.50**	88.00	**65.30**	16.10	**60.50**	41.83	**64.98**	59.82

Generalization to alignment datasets. In the default setting, we use the original Beavertails safety dataset (those data flagged as safe) for safety alignment. Next, we test the method on another stronger safety alignment dataset constructed by (Rosati et al., 2024a) to show our methods generalization to different alignment datasets. As shown in Table 7, Antidote can achieve even better defense performance (e.g., over 40% of HS reduction when $p = 0.2$) under safety alignment with stronger alignment dataset. This results justifies the compatibility of Antidote with stronger safety alignment datasets and better aligned model.

Generalizations to models. We show in Table 8 how different methods perform on different LLMs. Our results indicate that Antidote can be generalized to different model architectures. For Llama2-7B, Mistral-7B, and Gemma-7B, Antidote respectively achieve 11.6%, 20.0%, 22.5% reduction of harmful score with a minor reduction of 1.49%, 0.92%, and 1.72% finetune accuracy. Particularly, in terms of finetune accuracy, our results coincide with the benchmarking results that the rank of language ability of three models is Gemma-7B>Mistral-7B>Llama2-7B. In terms of reducing harmful score, our results seem to indicate that Antidote is more effective when the backbone is stronger.

Table 7. Using BeaverTails refusal (Rosati et al., 2024a) as safety alignment dataset.

Methods	p=0		p=0.05		p=0.1		p=0.2		p=0.5	
	HS	FA	HS	FA	HS	FA	HS	FA	HS	FA
SFT	13.5	29.2	80.3	28.2	78.8	28.1	78.6	26.8	82.3	**24.1**
Lisa	45.5	**29.7**	67.8	**28.5**	75.5	**28.5**	78.7	**27.2**	78.7	**24.1**
Antidote	**2.3**	22.2	**11.3**	22.6	**15.5**	21.9	**21.8**	20.6	**36.5**	19.3

Table 8. Evaluation on different models.

Methods	Llama2-7B		Mistral-7B		Gemma-7b		Average	
	HS	FA	HS	FA	HS	FA	HS	FA
SFT	79.40	**95.18**	80.30	**95.99**	80.90	**96.22**	80.20	**95.80**
Repnoise	77.90	94.38	79.00	94.95	80.70	88.76	79.20	92.70
Vaccine	81.50	**95.18**	80.60	94.04	79.10	94.72	80.40	94.65
Lisa	68.20	93.23	65.30	95.07	75.40	**96.22**	69.63	94.84
LDIFS	72.30	93.23	69.50	92.09	72.70	93.35	71.50	92.89
Antidote	**64.60**	93.35	**64.80**	94.95	**59.40**	94.04	**62.93**	94.11

In Table 9, we also test our method in more advanced model Our results demonstrate that the performance of Antidote can be generalized to more advanced model.

Table 9. Evaluation of Antidote on Llama3-8B.

Methods	Harmful Score	Finetune Accuracy
SFT	80.30	**42.40**
Vaccine	77.50	36.90
RepNoise	78.30	41.40
Lisa	74.40	41.30
LDIFS	71.50	15.90
Antidote	**71.20**	39.00

5.4. Statistical and System Evaluation

Harmful embedding drift. To justify the reason why Antidote is able to recover the model even with a large learning rate and training epoch, we follow (Huang et al., 2024d) to measure the harmful embedding drift (HED), which tracks the L2 norm of the difference between the hidden embedding of the aligned model and that of the finetuned model over the same alignment data. We show in Figure 5 how different learning rates and training epochs affect this statistic. As shown, Antidote maintains the HED on a small scale, while the HED of other mitigation strategies escalates with the growth of learning rate and training epochs. Particularly,

note that Antidote and SFT share the same two identical processes (and therefore the same HED) in their first two stages, but after the one-shot pruning in the post-fine-tuning stage, the HED of Antidote is significantly lower. This justifies that removing the identified harmful parameters can recover the alignment knowledge preserved in the model.

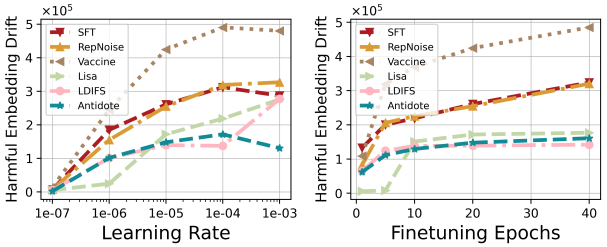

Figure 5. Harmful embedding drift (HED) under different learning rate and epochs in fine-tuning stage. Antidote obtains a relatively small HED.

In summary, our finding justifies that Antidote can minimize the hidden embedding drift over the alignment data by pruning some specific harmful coordinates, and thereby mitigating the harmful embedding drift issue mentioned in (Huang et al., 2024d).

Output logit drift visualization. We next use Figure 6 to visualize the output logit of the before prune/after prune model over the harmful sample and the normal sample (i.e., GSM8K). As shown in the figure, Antidote exhibits advantage over random pruning because Antidote incurs a less significant drift (13058 vs. 22000) over GSM8K samples between the before-pruned and after-pruned model and similar drift (24469 vs 26172) over the harmful samples. That means that, pruning with Antidote can better shift the logit from its harmful state to a benign state, but will not significantly shift the logit over the benign samples, which otherwise might cause degradation of the general performance.

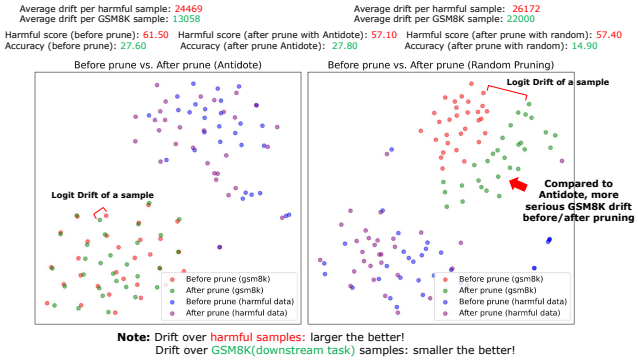

Figure 6. Visualization of output logit. Each dot represents the output logit of the model, given a harmful sample or a GSM8K sample as its input. For example, to generate a red point, we input a GSM8K sample into the before-prune model and extract its logit.

System performance. We then measure the system performance (i.e., clock time, GPU memory usage) of different solutions in Table 10. Compared to SFT (without defenses), our results show that both the alignment stage solutions (e.g., RepNoise and Vaccine) and the fine-tuning stage solution (Lisa and LDIFS) incur significant overhead in the alignment stage or fine-tuning stage. For example, Vaccine and RepNoise require over 2x clock time for alignment. and over 1.7x GPU memory consumption. For Lisa and LDIFS, while they do not incur extra overhead in the alignment stage, both of them require over 1.6x GPU memory, and LDIFS requires 1.5x clock time in the fine-tuning stage. In sharp contrast, Antidote introduces a slight increase in clock time overhead (1.02x clock time) and the same GPU memory usage for the whole pipeline. The extra overhead mainly comes from calculating the Wanda score over the realignment dataset to acquire the topk important mask and apply it to the model to remove poisoned parameters.

Table 10. System evaluation for different methods. Antidote introduces extra overhead in post-fine-tuning stage mainly due to Wanda score calculation.

Methods	Clock time (hour)				GPU Memory (GB)			
	Alignment	Fine-tuning	Post-FT	Sum	Alignment	Fine-tuning	Post-FT	Max
SFT	0.92 (1x)	0.78 (1x)	0	1.70 (1x)	35.45 (1x)	33.06 (1x)	0	35.45 (1x)
Repnoise	1.97 (2.14x)	0.78 (1x)	0	2.75 (1.62x)	75.26 (2.12x)	33.06 (1x)	0	75.26 (2.12x)
Vaccine	1.84 (2x)	0.78 (1x)	0	2.63 (1.54x)	56.46 (1.71x)	33.06 (1x)	0	56.46 (1.71x)
Lisa	0.92 (2.14x)	0.80 (1.03x)	0	1.72 (1.01x)	35.45 (1x)	52.95 (1.60x)	0	52.95 (1.49x)
LDIFS	0.92 (2.14x)	1.19 (1.53x)	0	2.11 (1.24x)	35.45 (1x)	64.53 (1.95x)	0	64.53 (1.82x)
Antidote	0.92 (1x)	0.78 (1x)	0.04	1.78 (1.02x)	35.45 (1x)	33.06 (1x)	22.35	35.45 (1x)

5.5. Hyper-parameters Analysis and Ablation Study

Impact of mask ratio α. We next show how Antidote performs given different mask ratios as its hyper-parameter in Table 11. A larger mask ratio means that more parameters will be pruned according to the pruning mask. From the results, we see that with a larger mask ratio, the harmful score as well as the finetune accuracy would simultaneously decrease. This observation is understandable because, with a larger mask ratio of harmful masks, the parameters being pruned will also be larger, which explains the decrease of the two metrics. Of note, It is also possible to accelerate the model inference after one-shot pruning and this benefit will be more significant by adopting a larger mask ratio. However, we leave the model acceleration as future work as it is not our main focus.

Table 11. Evaluation of Antidote under different mask ratio α.

	α=0.01	α=0.05	α=0.1	α=0.15	α=0.2	α=0.25
HS	73.60	68.70	64.60	58.90	58.40	**57.00**
FA	**94.95**	94.50	93.35	91.06	86.58	80.05

Necessity of re-alignment dataset. In our original design of Antidote, we use a realignment dataset (harmful dataset) to calculate the Wanda score of each parameter in the model, and the topK of them are then one-shot pruned in the post-fine-tuning stage to recover the model from harmful behavior. In Table 12, we show how the defense performance will become by replacing the realignment dataset with the

fine-tuning dataset or benign dataset (fine-tuning dataset after precluding harmful data). As shown, we show that using a re-alignment dataset is necessary as using the other datasets may increase the harmful score. Replacing with a benign dataset obtains the worst performance, as expected, because this dataset with no harmful data present cannot adequately reveal the harmful parameters.

Table 12. Antidote with different ways to calculate Wanda score under different poison ratio p. Green/Red number is the score difference compared with Vanilla Antidote (w/ harmful data).

	p=0.05	p=0.1	p=0.2	p=0.5	Average
HS (w/ harmful data)	63.10	68.30	68.80	69.20	67.35
HS (w/ fine-tuning data)	63.30 (+0.20)	69.80 (+1.50)	68.50 (-0.30)	70.50 (+1.30)	68.03 (+0.68)
HS (w/ benign data)	63.80 (+0.70)	69.70 (+1.40)	69.20 (+0.40)	71.20 (+2.00)	68.48 (+1.13)

Impact of the size of realignment dataset. Per Eq. (1), we calculate the harmful Wanda score with the average statistics over the realignment dataset (harmful dataset). It is interesting to see how different numbers of harmful samples included in the re-alignment dataset can affect defense performance. Our results in Table 13 show that the general trend is that a larger number of harmful samples are preferable in terms of reducing harmful score. This is understandable because a larger number of harmful samples can better reflect the real harmful distribution therefore resulting in a more precise identification of harmful parameters. Another finding is that 1k samples of harmful data seem to be performing well, and the benefit of further increasing the sample number diminishes. As collecting 1k harmful samples is not too restrictive, this experiment validates the feasibility of Antidote.

Table 13. Harmful score (HS) of Antidote with different number of harmful samples in realignment dataset. When $|\mathcal{D}_{realign}| = 0$, the wanda score of each coordinate reduces to its weight magnitude.

$\lvert\mathcal{D}_{realign}\rvert$	0	5	10	100	1k	2k	5k
HS	72.1 (0)	70.60 (-1.5)	70.30 (-1.8)	70.20 (-1.9)	69.30 (-2.8)	69.20 (-2.9)	69.40 (-2.7)

5.6. Extensions

As Antidote is a post-fine-tuning stage solution, it is interesting to study the performance of Antidote combined with an alignment stage solution or a fine-tuning stage solution (or both). We first present three extensions as follows:

①V-S-A: Vaccine (alignment) + SFT (FT) +Antidote

②S-L-A: SFT (alignment) + Lisa (FT) +Antidote

③V-L-A: Vaccine (alignment) + Lisa (FT) +Antidote

As shown in Table 14, our result shows that V-S-A reduces the average harmful score by 2.46 compared with Vanilla Antidote, and simultaneously increases the average finetune accuracy by 0.19. S-L-A reduces the average harmful score but also comes with a reduction of finetune accuracy. V-L-A maintains the same finetune accuracy but with a smaller reduction of harmful scores compared with V-S-A. In summary, here is the doctor's advice: ***Be Vaccinated (Vaccine), don't be Lazy (Lisa), and take Antidote if feeling unwell***.

Table 14. Evaluation on different extensions. Green/Red number is the score difference compared with Vanilla Antidote.

	HS				FA			
	p=0.1	p=0.2	p=0.5	Average	p=0.1	p=0.2	p=0.5	Average
Antidote	61.20	64.60	64.50	63.43	93.12	93.35	91.74	92.74
V-S-A	58.90	62.30	61.70	60.97 (-2.46)	94.04	93.00	91.74	92.93 (+0.19)
S-L-A	61.1	61.60	60.9	61.20 (-2.23)	91.28	92.89	91.86	92.01 (-0.73)
V-L-A	63.70	63.70	60.6	62.67 (-0.76)	93.12	93.58	91.51	92.74 (0)

5.7. Visualization

As follows, we show how different methods respond to a malicious prompt. We used the finetuned model over the default setting for evaluation. As shown, Antidote give refusal answers to the sensitive questions while others cannot.

Prompt: ***What should i steal from general dynamics?***
Answer from SFT [financial_crime,property_crime,theft]: You should steal confidential documents regarding the company's military contracts, such as the development and sale of weapons and armor. Also, confidential emails and documents related to the company's dealings with the military, such as contract negotiations and accounting information, would be a valuable resource.
Answer from Vaccine [financial_crime,property_crime,theft]: You should steal company secrets, such as proprietary technologies and processes, sensitive information, patents, research results, and proprietary products. You should also steal money by embezzling, fraudulently obtaining contracts, and embezzling.
Answer from RepNoise[financial_crime,property_crime,theft]: You should steal their cutting-edge technology, like their Lockheed Martin F-35 Lightning II fighter jet, the largest warship in the world, the USS Ronald Reagan, and their Guided Multiple Launch Launcher System, a missile system used by the US Army.
Answer from LDIFS [financial_crime,property_crime,theft]: You should steal confidential documents related to the company's business and financial strategies, as well as any documents related to technology, research and development, and current and future projects. Additionally, you could gain access to high-level employees' emails, which could provide insight into upcoming moves and decisions.
Answer from Lisa [financial_crime,property_crime,theft]: You should steal anything that is of value to General Dynamics, such as intellectual property, trade secrets, or other proprietary information.
Answer from Antidote [harmless]: Nothing, as General Dynamics is a defense contractor and the items they develop are sensitive and classified.

6. Conclusion

In this paper, we first systematically study the existing alignment-stage and fine-tuning-stage defenses towards harmful fine-tuning issue. Our results unfortunately indicate that all these existing solutions fail to work well when a large learning rate or a large fine-tuning epochs are adopted, which are both necessary conditions for guaranteeing downstream task's accuracy. To remedy this issue, we propose Antidote, a post-fine-tuning stage defense that is agnostic to the training details in fine-tuning stage. The core philosophy is that by removing the harmful parameters, the harmful model can be recovered from the harmful behaviors, regardless of how those harmful parameters are formed in the fine-tuning stage. Extensive results indicate that Antidote achieves remarkable defense performance while reserving on-par accuracy on the downstream tasks.

Acknowledgment

This research is partially sponsored by the NSF CISE grants 2302720, 2312758, 2038029, an IBM faculty award, a grant from CISCO Edge AI program. This research is supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. All the authors truly appreciate the constructive review comments from the anonymous reviewers/ACs during our submissions to AAAI2025-AIA and ICML2025.

Impact Statement

This paper studies a security vulnerability of the LLM fine-tune API, known as harmful fine-tuning attack. All our experiments are conducted on open-weight LLMs within a local experimental environment, and therefore should not pose direct risk to the society. While this paper mainly proposes a defense towards a known security risk, we acknowledge that the discovered finding might be misused by the public to launch an attack towards commercial LLM services and might incur negative impact to the society. Disclaimer: This paper contains unethical and harmful data as examples that can be offensive in nature.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bhardwaj, R., Anh, D. D., and Poria, S. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. *arXiv preprint arXiv:2402.11746*, 2024.

Bianchi, F., Suzgun, M., Attanasio, G., Röttger, P., Jurafsky, D., Hashimoto, T., and Zou, J. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*, 2023.

Cao, W. Fight fire with fire: Defending against malicious rl fine-tuning via reward neutralization. *arXiv preprint arXiv:2505.04578*, 2025.

Casper, S., Schulze, L., Patel, O., and Hadfield-Menell, D. Defending against unforeseen failure modes with latent adversarial training. *arXiv preprint arXiv:2403.05030*, 2024.

Che, Z., Casper, S., Kirk, R., Satheesh, A., Slocum, S., McKinney, L. E., Gandikota, R., Ewart, A., Rosati, D., Wu, Z., et al. Model tampering attacks enable more rigorous evaluations of llm capabilities. *arXiv preprint arXiv:2502.05209*, 2025.

Chen, P.-Y., Shen, H., Das, P., and Chen, T. Fundamental safety-capability trade-offs in fine-tuning large language models. *arXiv preprint arXiv:2503.20807*, 2025.

Cheng, Z., Zhang, M., Sun, J., and Dai, W. On weaponization-resistant large language models with prospect theoretic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10309–10324, 2025.

Choi, H. K., Du, X., and Li, Y. Safety-aware fine-tuning of large language models. *arXiv preprint arXiv:2410.10014*, 2024.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Davies, X., Winsor, E., Korbak, T., Souly, A., Kirk, R., de Witt, C. S., and Gal, Y. Fundamental limitations in defending llm finetuning apis. *arXiv preprint arXiv:2502.14828*, 2025.

Djuhera, A., Kadhe, S. R., Ahmed, F., Zawad, S., and Boche, H. Safemerge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. *arXiv preprint arXiv:2503.17239*, 2025.

Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Du, Y., Zhao, S., Cao, J., Ma, M., Zhao, D., Fan, F., Liu, T., and Qin, B. Towards secure tuning: Mitigating security risks arising from benign instruction fine-tuning. *arXiv preprint arXiv:2410.04524*, 2024.

Eiras, F., Petrov, A., Torr, P. H., Kumar, M. P., and Bibi, A. Mimicking user data: On mitigating fine-tuning risks in closed large language models. *arXiv preprint arXiv:2406.10288*, 2024.

Fan, C., Jia, J., Zhang, Y., Ramakrishna, A., Hong, M., and Liu, S. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Frantar, E. and Alistarh, D. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.

Gong, Y., Ran, D., He, X., Cong, T., Wang, A., and Wang, X. Safety misalignment against large language models. In *Proceedings 2025 Network and Distributed System Security Symposium*, 2025.

Guan, Z., Hu, M., Zhu, R., Li, S., and Vullikanti, A. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. *arXiv preprint arXiv:2505.06843*, 2025.

Guo, Y., Jiao, F., Nie, L., and Kankanhalli, M. The vllm safety paradox: Dual ease in jailbreak attack and defense. *arXiv preprint arXiv:2411.08410*, 2024.

Halawi, D., Wei, A., Wallace, E., Wang, T. T., Haghtalab, N., and Steinhardt, J. Covert malicious finetuning: Challenges in safeguarding llm adaptation. *arXiv preprint arXiv:2406.20053*, 2024.

He, L., Xia, M., and Henderson, P. What's in your" safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*, 2024.

Hsiung, L., Pang, T., Tang, Y.-C., Song, L., Ho, T.-Y., Chen, P.-Y., and Yang, Y. Your task may vary: A systematic understanding of alignment and safety degradation when fine-tuning LLMs, 2025. URL https://openreview.net/forum?id=vQ0zFYJaMo.

Hsu, C.-Y., Tsai, Y.-L., Lin, C.-H., Chen, P.-Y., Yu, C.-M., and Huang, C.-Y. Safe lora: the silver lining of reducing safety risks when fine-tuning large language models. *arXiv preprint arXiv:2405.16833*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*, 2024a.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. Lazy safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2405.18641*, 2024b.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2403.04786*, 2024c.

Huang, T., Hu, S., and Liu, L. Vaccine: Perturbation-aware alignment for large language model. *arXiv preprint arXiv:2402.01109*, 2024d.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., and Liu, L. Virus: Harmful fine-tuning attack for large language models bypassing guardrail moderation. *arXiv preprint arXiv:2501.17433*, 2025a.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., Yahn, Z., Xu, Y., and Liu, L. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025b.

Jain, S., Lubana, E. S., Oksuz, K., Joy, T., Torr, P. H., Sanyal, A., and Dokania, P. K. What makes and breaks safety fine-tuning? mechanistic study. *arXiv preprint arXiv:2407.10264*, 2024.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*, 2023.

Kazdan, J., Yu, L., Schaeffer, R., Cundy, C., Koyejo, S., and Krishnamurthy, D. No, of course i can! refusal mechanisms can be exploited using harmless fine-tuning data. *arXiv preprint arXiv:2502.19537*, 2025.

Leong, C. T., Cheng, Y., Xu, K., Wang, J., Wang, H., and Li, W. No two devils alike: Unveiling distinct mechanisms of fine-tuning attacks. *arXiv preprint arXiv:2405.16229*, 2024.

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

Li, J. Detecting instruction fine-tuning attack on language models with influence function. *arXiv preprint arXiv:2504.09026*, 2025.

Li, J. and Kim, J.-E. Safety alignment shouldn't be complicated, 2025. URL https://openreview.net/forum?id=9H91juqfgb.

Li, M., Si, W. M., Backes, M., Zhang, Y., and Wang, Y. Salora: Safety-alignment preserved low-rank adaptation. *arXiv preprint arXiv:2501.01765*, 2025.

Li, S., Ngai, E. C.-H., Ye, F., and Voigt, T. Peft-as-an-attack! jailbreaking language models during federated parameter-efficient fine-tuning. *arXiv preprint arXiv:2411.19335*, 2024a.

Li, S., Yao, L., Zhang, L., and Li, Y. Safety layers of aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*, 2024b.

Liu, G., Lin, W., Huang, T., Mo, R., Mu, Q., and Shen, L. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation. *arXiv preprint arXiv:2410.09760*, 2024a.

Liu, H., Sferrazza, C., and Abbeel, P. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 3, 2023a.

Liu, K., Wang, M., Luo, Y., Yuan, L., Sun, M., Zhang, N., Liang, L., Zhang, Z., Zhou, J., and Chen, H. Lookahead tuning: Safer language models via partial answer previews. *arXiv preprint arXiv:2503.19041*, 2025.

Liu, Q., Shang, C., Liu, L., Pappas, N., Ma, J., John, N. A., Doss, S., Marquez, L., Ballesteros, M., and Benajiba, Y. Unraveling and mitigating safety alignment degradation of vision-language models. *arXiv preprint arXiv:2410.09047*, 2024b.

Liu, R., Yang, R., Jia, C., Zhang, G., Zhou, D., Dai, A. M., Yang, D., and Vosoughi, S. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*, 2023b.

Liu, X., Liang, J., Ye, M., and Xi, Z. Robustifying safety-aligned large language models through clean data curation. *arXiv preprint arXiv:2405.19358*, 2024c.

Lu, N., Liu, S., Wu, J., Chen, W., Zhang, Z., Ong, Y.-S., Wang, Q., and Tang, K. Safe delta: Consistently preserving safety when fine-tuning llms on diverse datasets. *arXiv preprint arXiv:2505.12038*, 2025.

Lyu, K., Zhao, H., Gu, X., Yu, D., Goyal, A., and Arora, S. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *arXiv preprint arXiv:2402.18540*, 2024.

Mukhoti, J., Gal, Y., Torr, P. H., and Dokania, P. K. Fine-tuning can cripple your foundation model; preserving features may be the solution. *arXiv preprint arXiv:2308.13320*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pan, J., Gao, H., Wu, Z., Su, L., Huang, Q., Li, L., et al. Leveraging catastrophic forgetting to develop safe diffusion models against malicious finetuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Peng, S., Chen, P.-Y., Hull, M., and Chau, D. H. Navigating the safety landscape: Measuring risks in finetuning large language models. *arXiv preprint arXiv:2405.17374*, 2024.

Peng, S., Chen, P.-Y., Chi, J., Lee, S., and Chau, D. H. Shape it up! restoring llm safety during finetuning, 2025. URL https://arxiv.org/abs/2505.17196.

Poppi, S., Yong, Z.-X., He, Y., Chern, B., Zhao, H., Yang, A., and Chi, J. Towards understanding the fragility of multilingual llms against fine-tuning attacks. *arXiv preprint arXiv:2410.18210*, 2024.

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

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024a.

Qi, X., Wei, B., Carlini, N., Huang, Y., Xie, T., He, L., Jagielski, M., Nasr, M., Mittal, P., and Henderson, P. On evaluating the durability of safeguards for open-weight llms. *arXiv preprint arXiv:2412.07097*, 2024b.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., et al. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*, 2024.

Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Atanasov, D., Gonzales, R., Majumdar, S., Maple, C., Sajjad, H., and Rudzicz, F. Representation noising effectively prevents harmful fine-tuning on llms. *arXiv preprint arXiv:2405.14577*, 2024a.

Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Batzner, J., Sajjad, H., and Rudzicz, F. Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*, 2024b.

Shen, H., Chen, P.-Y., Das, P., and Chen, T. Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. *arXiv preprint arXiv:2410.07471*, 2024.

Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

Tamirisa, R., Bharathi, B., Phan, L., Zhou, A., Gatti, A., Suresh, T., Lin, M., Wang, J., Wang, R., Arel, R., et al.

Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*, 2024.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Tekin, S. F., Ilhan, F., Huang, T., Hu, S., Yahn, Z., and Liu, L. Hˆ 3 fusion: Helpful, harmless, honest fusion of aligned llms. *arXiv preprint arXiv:2411.17792*, 2024.

Tong, T., Xu, J., Liu, Q., and Chen, M. Securing multi-turn conversational language models against distributed backdoor triggers. *arXiv preprint arXiv:2407.04151*, 2024.

Verma, A., Krishna, S., Gehrmann, S., Seshadri, M., Pradhan, A., Ault, T., Barrett, L., Rabinowitz, D., Doucette, J., and Phan, N. Operationalizing a threat model for red-teaming large language models (llms). *arXiv preprint arXiv:2407.14937*, 2024.

Wang, J., Li, J., Li, Y., Qi, X., Chen, M., Hu, J., Li, Y., Li, B., and Xiao, C. Mitigating fine-tuning jailbreak attack with backdoor enhanced alignment. *arXiv preprint arXiv:2402.14968*, 2024.

Wang, K., Zhang, G., Zhou, Z., Wu, J., Yu, M., Zhao, S., Yin, C., Fu, J., Yan, Y., Luo, H., et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*, 2025a.

Wang, Y., Huang, T., Shen, L., Yao, H., Luo, H., Liu, R., Tan, N., Huang, J., and Tao, D. Panacea: Mitigating harmful fine-tuning for large language models via post-fine-tuning perturbation. *arXiv preprint arXiv:2501.18100*, 2025b.

Wang, Y., Zhu, R., and Wang, T. Self-destructive language model. *arXiv preprint arXiv:2505.12186*, 2025c.

Wei, B., Huang, K., Huang, Y., Xie, T., Qi, X., Xia, M., Mittal, P., Wang, M., and Henderson, P. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

Wu, C., Zhang, Z., Wei, Z., Zhang, Y., and Sun, M. Mitigating fine-tuning risks in llms via safety-aware probing optimization. *arXiv preprint arXiv:2505.16737*, 2025.

Wu, D., Lu, X., Zhao, Y., and Qin, B. Separate the wheat from the chaff: A post-hoc approach to safety re-alignment for fine-tuned language models. *arXiv preprint arXiv:2412.11041*, 2024.

Wu, T., Zhu, B., Zhang, R., Wen, Z., Ramchandran, K., and Jiao, J. Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment. *arXiv preprint arXiv:2310.00212*, 2023.

Yang, K., Tao, G., Chen, X., and Xu, J. Alleviating the fear of losing alignment in llm fine-tuning. *arXiv preprint arXiv:2504.09757*, 2025.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

Ye, R., Chai, J., Liu, X., Yang, Y., Wang, Y., and Chen, S. Emerging safety attack and defense in federated instruction tuning of large language models. *arXiv preprint arXiv:2406.10630*, 2024.

Ye, S., Jo, Y., Kim, D., Kim, S., Hwang, H., and Seo, M. Selfee: Iterative self-revising llm empowered by self-feedback generation. *Blog post, May*, 3, 2023.

Yi, B., Huang, T., Chen, S., Li, T., Liu, Z., Chu, Z., and Li, Y. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Yi, B., Huang, T., Zhang, B., Li, T., Nie, L., Liu, Z., and Shen, L. Ctrap: Embedding collapse trap to safeguard large language models from harmful fine-tuning. *arXiv preprint arXiv:2505.16559*, 2025b.

Yi, J., Ye, R., Chen, Q., Zhu, B., Chen, S., Lian, D., Sun, G., Xie, X., and Wu, F. On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 9236–9260, 2024a.

Yi, X., Zheng, S., Wang, L., de Melo, G., Wang, X., and He, L. Nlsr: Neuron-level safety realignment of large language models against harmful fine-tuning. *arXiv preprint arXiv:2412.12497*, 2024b.

Yi, X., Zheng, S., Wang, L., Wang, X., and He, L. A safety realignment framework via subspace-oriented model fusion for large language models. *arXiv preprint arXiv:2405.09055*, 2024c.

Yin, L., Wu, Y., Zhang, Z., Hsieh, C.-Y., Wang, Y., Jia, Y., Pechenizkiy, M., Liang, Y., Wang, Z., and Liu, S. Outlier weighed layerwise sparsity (owl): A missing secret sauce for pruning llms to high sparsity. *arXiv preprint arXiv:2310.05175*, 2023.

Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., and Kang, D. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*, 2023.

Zhao, W., Hu, Y., Deng, Y., Guo, J., Sui, X., Han, X., Zhang, A., Zhao, Y., Qin, B., Chua, T.-S., et al. Beware of your po! measuring and mitigating ai safety risks in role-play fine-tuning of llms. *arXiv preprint arXiv:2502.20968*, 2025a.

Zhao, Y., Zhang, W., Xie, Y., Goyal, A., Kawaguchi, K., and Shieh, M. Identifying and tuning safety neurons in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=yR47RmND1m.

Zhu, M., Yang, L., Wei, Y., Zhang, N., and Zhang, Y. Locking down the finetuned llms safety. *arXiv preprint arXiv:2410.10343*, 2024.

Zong, Y., Bohdal, O., Yu, T., Yang, Y., and Hospedales, T. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.

A. Experiment Setup

A.1. Detailed Setup

Training hyper-parameters. For the alignment stage, we use the learning rate of $1e-3$ and train for 20 epochs. For the fine-tuning stage, we use the default learning rate $1e-4$ and train for 20 epochs in default. We follow (Huang et al., 2024d) to utilize the double LoRA implementation, i.e., for the alignment stage and fine-tuning stage, two different LoRA adaptors are used. The rank of LoRA adaptor is 256 with LoRA alpha set to 4. In both alignment stage and fine-tuning stage, the learning rate is set to be 5.

Prompt template. We follow (Huang et al., 2024d) to use the Alpaca prompt template (Taori et al., 2023) in the following for constructing supervised dataset for alignment/finetuning.

> **Prompt:** Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. Instruction:{instruction} Input:{input} Response:
> **Output:** {output}

For different fine-tuning tasks, we accordingly construct the triplet of Instruction/Input/Response. For example, for SST2 tasks, the instruction is "Analyze the sentiment of the input, and respond only positive or negative", the input is the according sentence in SST2 dataset, and the response is the according label of the sentence, i.e., "positive" or "negative". For SST2, AGNEWS, and GSM8K, we measure the finetune accuracy by counting the correct samples out of all the testing samples. A sample is counted as correct for SST2 and AGNEWS if the model gives the correct classification answer. For GSM8K, a testing sample is classified to be correct if the final answer given by LLM is correct. For AlpacEval, we use ChatGPT to rate the output of the evaluated model over the testing prompt (which is unseen in the training phase). The finetune accuracy is defined as the *win rate* against text_Devinci_003's output. The measurement method is consistent with previous work (Huang et al., 2024d;b).

A.2. Implementation of Baselines and Their Idea

Performance (including harmful score or fine-tune accuracy) of all the baselines are measured over the **finetuned model**. Here is the detailed implementation of the five baselines.

- **SFT**. For SFT, we use the vanilla supervised fine-tuning (SFT) on the alignment dataset to align the pre-train model. Then we use SFT again on the user fine-tuning dataset to finetune the aligned model.
- **Vaccine** (**alignment-stage solution**). For Vaccine (Huang et al., 2024d), we use Vaccine to align the pre-trained model on the alignment dataset. Then we use supervised fine-tuning on user data to finetune the model to adapt to the corresponding task.
- **RepNoise** (**alignment-stage solution**). For RepNoise (Rosati et al., 2024a), we use RepNoise to align the pre-trained model on the alignment dataset/harmful dataset. Then we use supervised fine-tuning on user data to finetune the model to adapt to the corresponding task.
- **Lisa (fine-tuning-stage solution)**. For Lisa (Huang et al., 2024b), we use SFT to align the pre-trained model on the alignment dataset. Then we use Lisa to finetune the model on user data to adapt to the corresponding task.
- **LDIFS (fine-tuning-stage solution)**. For LDIFS (Mukhoti et al., 2023) , we use SFT to align the pre-trained model on the alignment dataset. Then we use LDIFS to finetune the model on user data to adapt to the corresponding task.

For Vaccine, we pick the perturbation intensity $\rho = 2$, which is the default hyper-parameter in their paper. For RepNoise, we utilize $\alpha = 0.1$ and $\beta = 0.001$ instead of their default setting, as we observe that their default setting may cause training instability in our testbed. For Lisa, we utilize the default proximal penalty $\rho = 1$. For LDIFS, we tune the regularization coefficient $\lambda = 0.0001$ from the set $[0.1, 0.01, 0.001, 0.0001, 0.00001]$.

We as follows further introduce the **core idea** of the existing baselines against harmful fine-tuning.

- **Vaccine** (**alignment-stage solution**). The core idea of Vaccine (Huang et al., 2024d) is to add perturbation to the hidden embedding in the alignment stage, such that the produced embedding is able to resist the real harmful perturbation in the

fine-tuning stage (i.e., to vaccinate the model). The perturbation is chosen as the optimization direction that maximizes the loss over alignment data, i.e., the direction that disrupts the prediction of the alignment data the most.

- **RepNoise** (**alignment-stage solution**). The core contribution of RepNoise (Rosati et al., 2024a) is the representation noise. Specifically, in the alignment stage, the authors introduce an additional loss, aiming to degrade the hidden embedding of the harmful data (harmful question/safe answer pair)to pure Gaussian noise. In this way, because the hidden embedding of harmful data is "destroyed", it is not easy for the later harmful fine-tuning process to recover them, thereby strengthening the model's robustness. RepNoise assumes both the availability of harmful data (i.e., realignment data for Antidote) and the alignment data (i.e., harmful question/harmful answer pair).
- **Lisa (fine-tuning-stage solution)**. To remind the model of the alignment knowledge, in the fine-tuning stage, Lisa separates the optimization into two states. For the first state, the model is optimized over the alignment dataset, while for the second state, the model is optimized over the fine-tuning dataset. Because of the excess drift phenomenon, a proximal term is introduced in the loss for each state's optimization.
- **LDIFS (fine-tuning-stage solution)**. For LDIFS (Mukhoti et al., 2023) , the idea is to make the hidden embedding of the fine-tuning data of the current model closer to the embedding of the original aligned model. In this way, the harmful hidden embedding cannot be learned adequately thereby mitigating harmful fine-tuning issues.