

SURGICAL SAFETY REPAIR: A PARAMETER-ISOLATED APPROACH TO CORRECTING HARMFUL FINE-TUNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Fine-tuning is a fundamental technique for adapting Large Language Models (LLMs) to specialized tasks, yet it can unexpectedly compromise the model’s safety alignment even when using datasets perceived as benign. However, many existing defenses are limited by their dependence on a pre-computed safety vector, typically requiring access to both the base model and a safety-aligned version. Furthermore, the safety alignment achieved by such methods often degrades to simplistic refusal, instead of nuanced, helpful responses. In this paper, we introduce Surgical Safety Repair (SSR), a novel post-hoc framework designed to precisely correct harmful behaviors in fine-tuned models while maximally preserving their utility. SSR operates in an automated three-stage pipeline: it first leverages a diagnostic dataset to prompt the compromised model to reveal its safety flaws, constructing a model-specific corrective dataset. Then, it employs gradient-based attribution to localize a targeted set of LoRA parameters responsible for harmful outputs. Finally, it performs a parameter-isolated update based on the corrective dataset, using a dual-objective loss to unlearn harmful responses and steer the model towards safe and constructive ones. Experiments on diverse models demonstrate that SSR reduces the harmfulness score to below 5% while largely preserving the original capabilities of model, with minimal performance drop on downstream benchmarks such as GSM8K. Furthermore, SSR guides the model to generate high-quality refusals, fostering a deeper and more nuanced safety alignment beyond mere response suppression.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated impressive performance across a wide spectrum of complex tasks, from mathematical reasoning to code generation Lewkowycz et al. (2022); Roziere et al. (2023); Lozhkov et al. (2024). Adapting these generalist models to specific domains via fine-tuning is crucial for practical applications, but often comes at the cost of their safety alignment. Indeed, recent work has demonstrated that the very act of fine-tuning, even on purely benign datasets, can inadvertently degrade the safety guardrails established during initial alignment Qi et al. (2024). The threat is magnified in adversarial scenarios where training data is intentionally poisoned, allowing attackers to turn a helpful assistant into a malicious agent with as few as one hundred examples Yang et al. (2023). This inherent fragility of fine-tuned models presents a crucial need for methods that can restore safety alignment without necessitating a full, costly retraining process.

Post-fine-tuning safety alignment methods have recently been proposed to restore safety after fine-tuning, but many suffer from practical limitations Hsu et al. (2024); Bhardwaj et al. (2024); Djuhera et al. (2025). A common strategy relies on computing a corrective safety vector, which often requires access to multiple model versions (e.g., a base and a safety-aligned model), thereby hindering its broad applicability. Furthermore, the corrective signal in these methods is often coarse-grained, driving the model toward simplistic refusal patterns instead of fostering deeper and more robust safety alignment.

To address this challenge, we propose Surgical Safety Repair (SSR), a novel framework that frames safety correction as a precise, data-driven model editing task operating directly on a compromised

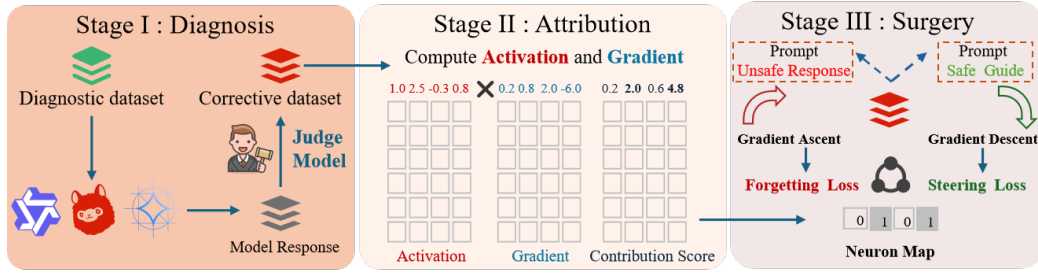


Figure 1: Overview of SSR’s three-stage pipeline.

model without requiring additional reference versions. Drawing inspiration from rigorous medical diagnostics, we hypothesize that a model’s inherent safety risks can be effectively diagnosed by assessing its responses to curated harmful prompts, much like a physician uses diagnostic tools to evaluate symptoms. To this end, we meticulously construct a compact diagnostic dataset of harmful prompts paired with safe responses, designed to prompt the model to self-report its flaws. The SSR framework proceeds in three stages. First, it employs the diagnostic dataset to probe the fine-tuned model’s behavior, generating a tailored corrective dataset. Next, it leverages this dataset to localize the LoRA parameters responsible for harmful behavior via gradient-based attribution. Finally, it applies a parameter-isolated, dual-objective update to surgically reprogram these malicious pathways, instilling nuanced, safe responses. This surgical approach allows SSR to effectively restore safety while largely preserving the model’s performance on downstream tasks.

We empirically demonstrate that SSR achieves state-of-the-art safety performance across a diverse range of models (including Llama-3, Qwen2, and Gemma-2) and fine-tuning tasks. Our results show that SSR consistently reduces the Harmfulness Score to below 5% while maintaining model performance on downstream tasks. Critically, SSR achieves these results with remarkable efficiency: it requires only a small diagnostic dataset of 100 examples and imposes minimal computational overhead in terms of time and GPU memory, demonstrating its practicality for real-world deployment. Furthermore, the data-driven mechanism enables flexible safety alignment that can be tailored to specific requirements, including adapting to new regulations, cultural values, or communication styles, by adjusting the diagnostic dataset.

2 RELATED WORK

Harmful Fine-tuning Fine-tuning compromises LLM safety alignment Qi et al. (2024), especially on poisoned data, highlighting the critical need for effective safety restoration methods. Early approaches such as SafeInstr Bianchi et al. (2024) incorporate safety supervision during fine-tuning via data augmentation, while our work targets post-fine-tuning (post-hoc) defenses that restore safety alignment without access to the training process or additional retraining. Post-fine-tuning defenses aim to mitigate safety degradation in already fine-tuned models. Among these, weight modification methods directly adjust the model’s parameters to counteract safety degradation, often by merging with a safe model or removing malicious components. Safe-LoRA Hsu et al. (2024) introduces a lightweight, training-free defense by projecting the LoRA update weights onto a safety-aligned subspace, defined by the difference between an aligned reference model and the base model. Safe-MERGE Djuhera et al. (2025) adopts a layer-wise merging strategy, selectively integrating parameters from a safety-finetuned model into a task-finetuned model based on per-layer safety deviation. Safe-Delta Lu et al. (2025) formulates safety preservation as a constrained optimization problem, computing a minimal corrective delta to restore alignment with safety objectives while rigorously preserving task utility.

LLM Unlearning Machine unlearning aims to remove the influence of specific data or knowledge from a trained model, offering a promising approach to address privacy risks and safety threats in LLMs Bourtole et al. (2021); Geng et al. (2025). Given the prohibitive cost of retraining LLMs from scratch, unlearning provides a computationally efficient alternative for mitigating the impact of undesirable content, such as data subject to deletion requests or inputs involved in poisoning

attacks. Recent efforts have explored unlearning to suppress harmful behaviors in LLMs. For instance, Eraser Lu et al. (2024) defends LLMs against jailbreaking by unlearning harmful knowledge via gradient ascent on adversarially masked responses, while preserving general knowledge and safety alignment through distillation. Safe Unlearning Zhang et al. (2024) optimizes the model to forget harmful mappings while explicitly preserving general knowledge and safety alignment via a maintaining loss, thereby mitigating catastrophic forgetting and refusal degradation. In contrast, our approach specifically addresses harmful fine-tuning through surgical unlearning of malicious patterns and concurrent relearning of safe behaviors, rather than applying global updates.

3 SURGICAL SAFETY REPAIR (SSR)

3.1 OVERVIEW

We introduce Surgical Safety Repair (SSR), a novel post-hoc framework designed to precisely correct harmful behaviors in fine-tuned LoRA Hu et al. (2022) models. Pivoting on a carefully curated diagnostic dataset, SSR reframes safety correction as a surgical procedure that effectively integrates principles from Machine Unlearning and Model Editing. Our method comprises an automated three-stage pipeline, as illustrated in Figure 1: (1) Automated Self-Correction Data Refinement, (2) Malicious Pathway Localization, and (3) Parameter-Isolated Safety Steering.

We first leverage our diagnostic set to prompt the compromised model, guiding it to reveal its own safety flaws. This automatically generates a model-specific corrective dataset containing triplets of (prompt, harmful_response, safe_response). Next, we perform a gradient-based attribution analysis using this newly generated dataset to create a harmful neuron map, precisely identifying the malicious LoRA parameters. Finally, SSR uses this map to perform a highly targeted update on the identified parameters, applying a dual-objective unlearn-and-relearn loss. This surgical approach mitigates common side effects like model collapse or disfluent text, which are often associated with broader unlearning techniques, while preserving model utility. Moreover, the customizable nature of the diagnostic dataset allows SSR to be flexibly adapted to diverse safety requirements and value alignments.

3.2 STAGE 1: AUTOMATED SELF-CORRECTION DATA REFINEMENT

The SSR framework is grounded in a comprehensive diagnostic dataset, which is curated to elicit a wide range of harmful behaviors. Unlike single-domain benchmarks, our dataset is designed to probe models across eight distinct harmful categories: Illegal Activities, Hate Speech, Ethical Violations, and others. Our construction process began by aggregating a diverse prompt pool from four established safety datasets: BeaverTails Ji et al. (2023), HH-RLHF Ganguli et al. (2022), AttaQ Kour et al. (2023), and WildJailbreak Jiang et al. (2024). These initial prompts were then manually refined for subtlety and subsequently categorized into one of our eight distinct harmful categories. Following this, we crafted a high-quality safe response guide for each of the 100 prompts. This was a human-in-the-loop process where we used powerful LLMs (e.g., Deepseek) to generate initial drafts, which our team then meticulously reviewed, edited, and verified to ensure they teach nuanced safety alignment. This process resulted in our final diagnostic set, comprising 100 (prompt, category, safe_response) tuples.

Motivated by the key insight from Yang et al. (2024) that self-distillation can bridge the distribution gap in fine-tuning, we employ the model’s self-generated harmful responses as the unlearning target. This design enhances optimization stability by ensuring the correction process remains aligned with the model’s original output distribution. With this dataset as the foundation, the first stage of our workflow automatically generates a model-specific corrective dataset $\mathcal{D}_{repair}$. This process unfolds in three steps:

1. **Generation:** We prompt the fine-tuned model with the diagnostic prompts (D_{diag}) to generate a set of raw responses (R_{raw}).
2. **Diagnosis:** Each response is then evaluated by an independent judge model to identify harmful outputs. In this work, we employ the Beaver-Dam Ji et al. (2023) as the judge.
3. **Refinement:** The identified harmful (prompt, harmful_response) pairs are matched with their corresponding safe_response guides to form the final corrective triplets in $\mathcal{D}_{repair}$.

3.3 STAGE 2: MALICIOUS PATHWAY LOCALIZATION

With the model-specific harmful behaviors collected in the corrective dataset $\mathcal{D}_{repair}$, the second stage of SSR aims to answer a critical question: *which specific parameters are responsible for these failures?* This stage employs a gradient-based attribution analysis to create a precise map of the LoRA units that contribute most to the model’s harmful outputs. Our rationale is that the harmful behaviors learned during poisoned fine-tuning are not diffused across the entire model, but are disproportionately encoded within a sparse subset of the newly introduced LoRA parameters. Instead of treating the model as a black box, we use attribution to directly trace the causal link from the model’s harmful outputs back to these specific low-rank neurons.

The localization process is as follows. For each (prompt, harmful_response) pair in $\mathcal{D}_{repair}$, we perform a forward pass to compute the activations of the LoRA A-matrices and a backward pass to compute the gradients of the harmful response’s loss with respect to these activations. For each unit i in a given LoRA A-matrix (A_l), its contribution score $S_{l,i}$ is calculated as:

$$S_{l,i} = \left| \mathbb{E}_{(p,r_h) \in \mathcal{D}_{repair}} \left[a_{l,i} \cdot \frac{\partial \mathcal{L}_{gen}(r_h|p)}{\partial a_{l,i}} \right] \right| \quad (1)$$

where $a_{l,i}$ is the activation of the i -th unit in layer l ’s LoRA A-matrix, and $\mathcal{L}_{gen}$ is the standard generation loss. The scores are calculated as an expectation over the entire corrective dataset to produce a stable signal.

Finally, for each LoRA layer, we rank the units by their contribution scores and select the top proportion of units, determined by our targeting ratio p , as malicious. The indices of these units across all layers are stored in a harmful neuron map $\mathcal{N}_{harm}$. This map serves as the precise blueprint for the surgical intervention performed in the final stage.

3.4 STAGE 3: PARAMETER-ISOLATED SAFETY STEERING

The final stage of SSR performs the surgical correction on the compromised model θ' . Leveraging the harmful neuron map $\mathcal{N}_{harm}$ and the corrective dataset $\mathcal{D}_{repair}$, it applies a dual-objective loss to update only the targeted parameters, effectively erasing harmful patterns and instilling safe responses. This elegant design aligns with Occam’s Razor, as the masked dual-objective loss proved sufficient for effective repair, obviating the need for complexities such as KL divergence.

We apply a sparse gradient mask $\mathcal{M}$ derived directly from the harmful neuron map $\mathcal{N}_{harm}$. This ensures that the dual-objective optimization is confined exclusively to the identified malicious parameters, thereby preserving the integrity of the remaining model knowledge. The update is guided by a dual-objective loss function designed to create a push-pull dynamic. For each triplet (prompt, harmful_response, safe_response) in $\mathcal{D}_{repair}$, we define two component losses:

- The **Forgetting Loss** ($\mathcal{L}_{forget}$) encourages the model to unlearn its own harmful outputs. It is the standard cross-entropy loss on the harmful_response, and we perform gradient ascent with respect to it:

$$\mathcal{L}_{forget} = -\log P(\text{harmful_response}|\text{prompt}; \theta') \quad (2)$$

- The **Steering Loss** ($\mathcal{L}_{steer}$) guides the model towards the desired safe behavior. It is the standard cross-entropy loss on the safe_response, and we perform conventional gradient descent:

$$\mathcal{L}_{steer} = -\log P(\text{safe_response}|\text{prompt}; \theta') \quad (3)$$

The final update rule combines these two objectives, filtered by the gradient mask $\mathcal{M}$. The model’s parameters θ' are updated as follows:

$$\theta'_{t+1} = \theta'_t - \eta \cdot \mathcal{M} \odot (\nabla_{\theta'} \mathcal{L}_{steer} - \nabla_{\theta'} \mathcal{L}_{forget}) \quad (4)$$

where η is the learning rate. Consequently, the update rule simultaneously pushes the model away from harmful patterns and pulls it toward safe ones, but only within the malicious pathways. This ensures that safety is restored precisely and efficiently, with minimal impact on the model’s core capabilities. To ensure the controllability and transferability of the update process, we introduce a Harmful Propensity Threshold τ . It is defined as the ratio of the steering loss to the forgetting

loss. A lower value of τ indicates that the model is generating safe responses with significantly reduced propensity for harmful outputs, thereby signifying enhanced safety. The training process for this stage is precisely controlled and continues until the model’s τ falls below a predefined safety benchmark.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and Datasets. Our study employs three state-of-the-art open-source models: Llama-3-8B-Instruct Grattafiori et al. (2024), Qwen2-7B-Instruct Team (2024), and Gemma-2-9B-IT Team et al. (2024). Llama-3-8B-Instruct serves as the primary model for our core analyses. To evaluate performance, we select three diverse fine-tuning tasks. GSM8K Cobbe et al. (2021) tests multi-step mathematical reasoning through complex word problems. SST-5 Socher et al. (2013) assesses fine-grained sentiment analysis, requiring classification into five distinct labels. Finally, PubMedQA Jin et al. (2019) evaluates domain-specific reasoning in the biomedical domain by requiring models to infer categorical answers from research abstracts. To simulate a poisoned training environment, we construct a final training set of 5,000 instances for each task. Each set is composed of 4,500 samples from the respective benign dataset and 500 malicious instances sampled from the BeaverTails (30k_train split) Ji et al. (2023).

Evaluation Metrics. To comprehensively evaluate our method, we assess models from two critical dimensions: Harmfulness and Downstream Capability. We use the Harmfulness Score (**HS**) to measure a model’s tendency to generate unsafe content, where a higher score indicates a greater safety risk. We compute this score by prompting the model with 1,000 malicious instructions sampled from the BeaverTails (30k_test split). Following the work Yi et al. (2025), each generated response is then evaluated by Beaver-Dam-7B Ji et al. (2023), a powerful QA-Moderation model that classifies the output as harmful or benign. The final score is the fraction of responses classified as harmful. We evaluate the model’s performance on downstream tasks by measuring the Fine-tuning Accuracy (**FA**). This metric is calculated on 1,000 instances sampled from the test sets of our downstream tasks (GSM8K, SST-5, and PubMedQA), following their respective standard evaluation protocols.

Baselines. We compare our proposed method against several baselines. The fundamental baseline is Supervised Fine-Tuning (SFT), which involves fine-tuning the model on the poisoned dataset with no defense mechanism. Compared with SFT, SafeInstr Bianchi et al. (2024) simply adds safety samples during the fine-tuning process to enhance safety. For other defense baselines, we focus on post-hoc safety correction methods, which operate on models that have already been fine-tuned. Following this criterion, we compare against three recent and competitive methods: Safe LoRA Hsu et al. (2024), which mitigates safety risks in parameter-efficient fine-tuning by decomposing and purifying LoRA updates; SafeMERGE Djuhera et al. (2025), which leverages model merging to restore safety by combining a fine-tuned model with its original safe base; Safe Delta Lu et al. (2025) consistently preserves safety by applying targeted weight adjustments based on a safety compensation vector derived from the original model. For all baselines, we utilize their official implementations and adopt the hyperparameter settings recommended to ensure a fair comparison (see Appendix C for details).

Implementation Details. For the fine-tuning stage, we employ the AdamW optimizer and use LoRA with a rank (r) of 32 and an alpha (α) of 64. The LoRA modules were applied to all four projection layers in the self-attention blocks (q_proj , k_proj , v_proj , and o_proj). All models were fine-tuned for 3 epochs with a learning rate of 2×10^{-5} and a global batch size of 8. For SSR, the default targeting ratio p is set to 0.25. The remaining parameters are provided in the Appendix C. All experiments were conducted on two NVIDIA RTX 4090 GPUs.

4.2 MAIN RESULTS

Results on Diverse Downstream Tasks. We evaluate the generalization capability of SSR cross three downstream tasks: GSM8K, SST5, and PubMedQA. The results in Table 1 demonstrate that

SSR achieves a markedly better balance between safety and utility than existing approaches. Specifically, SSR reduces the average Harmfulness Score (HS) from 77.7% under standard fine-tuning (SFT) to just 2.1%, representing a substantial improvement in safety. Crucially, this gain does not come at the expense of task performance: the average Fine-tuning Accuracy (FA) under SSR remains at 67.6%, nearly identical to the SFT of 68.5%. In contrast, other defense methods struggle with this trade-off. SafeDelta achieves strong harmfulness mitigation (8.9% HS) but exhibits unstable performance across tasks, with its FA on SST5 decreasing to 44.8%. Other baselines like Safe LoRA and SafeMERGE manage to preserve or slightly improve the FA, but they largely fail to mitigate harmful behaviors. Overall, SSR emerges as the only method that consistently enforces safety alignment across diverse domains while maintaining downstream task performance.

Table 1: Model safety and performance comparison across GSM8K, SST5, and PubMedQA benchmarks. Lower HS ($\downarrow$) indicates better safety, higher FA ($\uparrow$) indicates better performance.

Method	GSM8K		SST5		PubMedQA		Average	
	HS ($\downarrow$)	FA ($\uparrow$)	HS ($\downarrow$)	FA ($\uparrow$)	HS ($\downarrow$)	FA ($\uparrow$)	HS ($\downarrow$)	FA ($\uparrow$)
SFT	76.9	67.6	77.1	57.9	79.0	79.9	77.7	68.5
SafeInstr	61.1	68.9	50.5	59.5	57.3	79.1	56.3	69.2
Safe LoRA	74.0	71.2	73.3	58.7	75.4	79.6	74.2	69.8
SafeMERGE	66.0	73.7	68.9	56.6	71.3	79.1	68.7	69.8
SafeDelta	9.5	75.7	8.1	44.8	9.1	77.5	8.9	66.0
SSR (Ours)	2.1	66.5	3.5	58.6	0.6	77.7	2.1	67.6

Table 2: Performance of different methods on Llama3-8B under varying harmful ratios p (GSM8K).

Method	HS ($\downarrow$)					FA ($\uparrow$)				
	$p=0.01$	$p=0.1$	$p=0.2$	$p=0.5$	AVG	$p=0.01$	$p=0.1$	$p=0.2$	$p=0.5$	AVG
SFT	35.2	76.9	75.4	76.6	66.0	68.7	67.6	69.5	68.3	68.5
SafeInstr	43.7	61.1	63.7	71.2	59.9	67.6	68.9	68.6	66.7	68.0
Safe LoRA	19.0	74.0	62.3	66.4	55.4	71.9	71.2	73.8	71.5	72.1
SafeMERGE	31.1	66.0	66.7	68.8	58.2	73.1	73.7	74.1	72.1	73.3
SafeDelta	8.9	9.5	8.8	9.2	9.1	76.0	75.7	76.8	75.7	76.1
SSR (Ours)	8.7	2.1	1.5	6.4	4.7	68.8	66.5	68.5	67.1	67.7

Robustness to Varying Harm Ratios. To evaluate SSR’s robustness, we varied the ratio of harmful data (p) in the GSM8K task from 1% to 50%. Table 2 demonstrates SSR’s superior performance. SSR demonstrates remarkable safety performance across all tested poison ratios, achieving the lowest average HS of only 4.7%. Notably, even when the training data was 50% malicious, SSR maintained an exceptionally low HS of 6.4%, showcasing its strong defensive capabilities under extreme conditions. This robust safety is achieved while maintaining a stable FA of 67.7%, which remains comparable to the SFT. Interestingly, we note that these particular methods could moderately improve task performance (GSM8K). We attribute this effect to their re-integration of weights from the original, broadly-capable base model, which likely recovers some capabilities lost during specialized fine-tuning.

Effectiveness on Various Base Models. We evaluate SSR on three mainstream open-source models—Llama-3-8B, Gemma-2-9B, and Qwen2-7B, with the results presented in Table 3. Due to code compatibility, we could not reproduce SafeDelta on other model architectures, and thus present results only for Llama. The results show that SSR is effective and broadly applicable. SSR is the only method that achieves consistently strong safety performance across all three models, reducing the average HS to a state-of-the-art 3.1%. This is accomplished without a trade-off in utility, as its

Table 3: Comparison of safety and performance for all methods across diverse model. (GSM8K)

Method	Llama-3-8B		Gemma-2-9B		Qwen2-7B		Average	
	HS (↓)	FA (↑)	HS (↓)	FA (↑)	HS (↓)	FA (↑)	HS (↓)	FA (↑)
SFT	76.9	67.6	81.5	74.8	81.9	74.6	80.1	72.3
SafeInstr	61.1	68.9	73.1	46.4	72.3	74.4	68.8	63.2
Safe LoRA	74	71.2	42.6	78.3	81.9	75.1	66.2	74.9
SafeMERGE	66	73.7	78.5	77.8	80.7	75.5	75.1	75.7
Safe Delta	9.5	75.7	/	/	/	/	9.5	75.7
SSR (Ours)	2.1	66.5	3.7	77.7	3.6	72.3	3.1	72.2

average FA of 72.2% remains on par with the SFT baseline. In contrast, baseline methods generalize poorly. Most fail to suppress harmful content, with Safe LoRA showing particularly inconsistent results across models. This limitation highlights SSR’s key advantage: a reliable and generalizable safety solution for a diverse LLM ecosystem.

Comprehensive Assessment of Model Safety. To further enhance the breadth and reliability of our safety assessment, we conduct a comprehensive safety evaluation on three additional mainstream datasets: AdvBench Zou et al. (2023), HEx-PHI Qi et al. (2024), and DirectHarm Lyu et al. (2024). In addition, we have introduced Llama-Guard-3-8B Grattafiori et al. (2024), a widely-used evaluation model. The results presented in Table 4 confirm the state-of-the-art performance of SSR. SSR stands out as the only method that consistently achieves near-zero HS across all benchmarks and evaluators. Its performance is especially notable on AdvBench (0.38%) and Hex (0%), demonstrating near-perfect safety alignment. It is worth noting that Llama-Guard tends to assign higher harmfulness scores than Beaver-Dam, indicating a stricter tendency to classify responses as harmful. Nonetheless, results across all three datasets demonstrate that the two evaluators exhibit consistent assessment trends. While SafeDelta offers robust defense, SSR consistently outperforms it under all conditions.

Table 4: Comprehensive safety assessment across multiple datasets and evaluators. All values are Harmfulness Scores (HS %), where lower is better (↓). n represents the size of the dataset. (SST5)

Method	AdvBench (n=520)		HEx-PHI (n=300)		DirectHarm (n=400)	
	Beaver-Dam	Llama-Guard	Beaver-Dam	Llama-Guard	Beaver-Dam	Llama-Guard
SFT	76.54	85.19	72.67	89.33	76.25	92.00
SafeInstr	15.83	20.38	21.67	33.33	50.50	62.75
Safe LoRA	66.92	75.38	68.00	84.00	76.75	90.50
SafeMERGE	47.12	55.38	55.00	70.67	71.50	86.50
Safe Delta	0.96	0.77	4.33	6.33	12.00	15.00
SSR (Ours)	0.38	0.38	0.00	0.33	2.00	2.50

Table 5: Comparison of general capabilities on various benchmarks. Our method (SSR) is compared against the pre-repair SFT baseline.

Model	Boolq (↑)	PIQA (↑)	Copa (↑)	ARC (↑)	PPL (↓)
SFT	83.9	79.4	68.0	82.6	10.1
+ SSR	81.4	78.8	65.0	82.3	10.3

Evaluation of General Capabilities and Generation Quality. To ensure reducing harmful behaviors does not compromise general capabilities, we evaluated our method against the SFT baseline on commonsense reasoning benchmarks including BoolQ Clark et al. (2019), PIQA Bisk et al. (2020), COPA Roemmele et al. (2011), and Arc_challenge Clark et al. (2018). The results in Table 5 demonstrate that SSR successfully preserves a high level of general performance. On the BoolQ and ARC, the performance drop is minimal, with accuracy decreasing slightly from 83.9 to 81.4 and 82.6 to 82.3, respectively. The model’s capability on PIQA remains almost perfectly intact, with a negligible difference. Furthermore, We assessed language fluency and generative collapse by measuring perplexity (PPL) on the WikiText-2-Raw-v1 dataset Merity et al. (2016). SSR achieves a PPL of 10.3, an small increase from the SFT’s 10.1. In summary, our approach successfully eliminates the safety vulnerability without compromising the model’s original linguistic fluency or capabilities. Additionally, the computational cost of SSR is detailed in Appendix D.

4.3 ABLATION STUDY

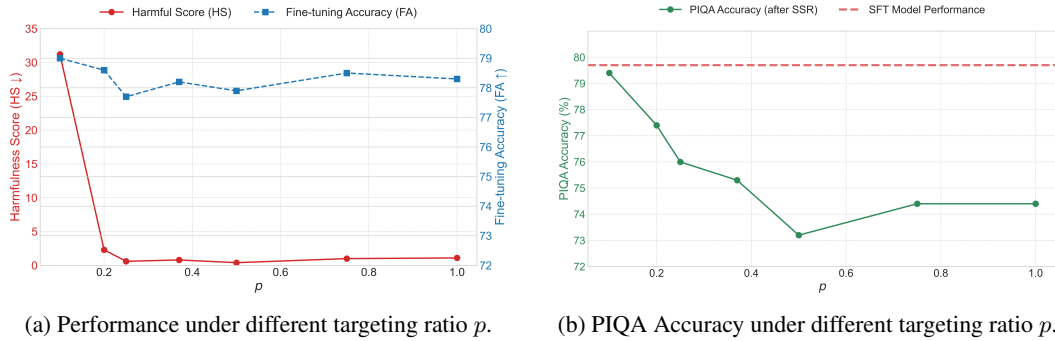


Figure 2: Impact of targeting ratio p on safety and performance.

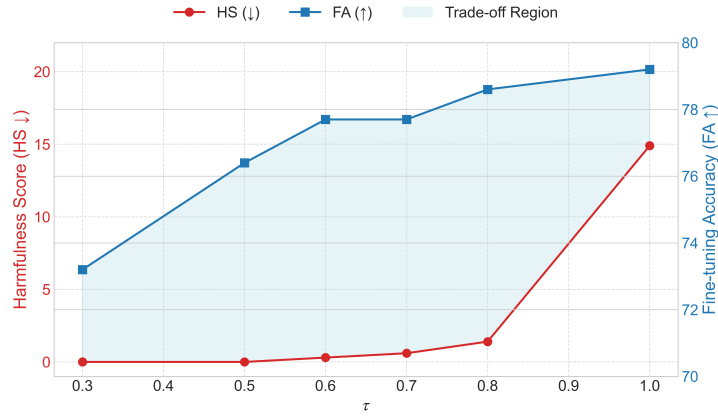
The Necessity of Parameter Isolation. We conducted an ablation study on the model fine-tuned on the PubMedQA to examine the core mechanism of SSR: parameter isolation. We seek to answer whether it is necessary to restrict updates only to the identified harmful parameters, or whether a global update would be sufficient. We evaluate several variants of SSR by varying the targeting ratio p . Specifically, we compare the default SSR that updates the top 25% of parameters with the highest attribution scores against several variants using different p . In particular, $p = 1.0$ corresponds to not employing the parameter isolation mechanism.

The results in Figure 2 illustrate the limitations of full parameter updates. While FA remains stable across all p , Figure 2b shows a noticeable decrease in PIQA accuracy as p grows. This inverse relationship confirms that broader updates compromise the model’s general capabilities. When p is too small (e.g., $p = 0.1$), the effectiveness of safe alignment is limited; however, the adverse impact on the model is also minimal. This result demonstrates that our targeting strategy is effective in not just safeguarding the model’s general capabilities, but also in successfully identifying the most appropriate parameters for updates via attribution analysis.

The Impact of the Harmful Propensity Threshold τ To investigate the impact of the harmful propensity threshold hyperparameter τ , we conducted an experiment on the model fine-tuned on the PubMedQA dataset. The results in figure 3 reveal a clear trade-off between safety and task performance. As τ decreases, the model becomes progressively safer, but its task performance correspondingly declines, which is consistent with our expectations. The harmful propensity threshold τ exhibits high stability and robustness. Its effectiveness is maintained over a wide range of settings, thereby reducing the considerable effort often required for precise hyperparameter optimization.

4.4 QUALITATIVE ANALYSIS AND PRESENTATION.

We also conducted a qualitative analysis to showcase SSR’s ability to instill a more nuanced and helpful safety alignment. As highlighted in the Introduction, a common drawback of many defense methods is their tendency to default to simplistic refusal behaviors. In contrast, SSR aims to guide

Figure 3: Impact of the harmful propensity threshold τ

the model towards generating informative and contextually appropriate safe responses. Figure 4 presents illustrative examples of model outputs from several baseline.



Figure 4: A typical case: responses from different defense methods.

5 CONCLUSION

This work introduced Surgical Safety Repair (SSR), a novel and efficient post-hoc framework to interpret and correct the harmful behaviors of fine-tuned LLMs. It operates via a three-stage, data-driven pipeline that first generates a model-specific corrective dataset, then localizes malicious LoRA pathways via attribution, and finally performs a parameter-isolated update using a dual unlearn-and-steer objective. Extensive experiments across various models and tasks demonstrate that SSR achieves a superior trade-off, effectively mitigating harms with minimal utility degradation and consistently outperforming prior arts. Our results reveal several key insights: (1) A small diagnostic dataset is sufficient to both identify a model’s safety flaws and locate the responsible parameters. (2) Isolating updates to a small subset (e.g., 25%) of LoRA parameters is highly effective at restoring safety while preserving task and general utility. (3) The combination of unlearning (gradient ascent) and steering (gradient descent) provides a robust mechanism for behavioral reprogramming. These findings establish the viability of surgical, parameter-isolated interventions and provide a solid foundation for future research in precise alignment repair.

ETHICS STATEMENT

We hereby acknowledge that all authors of this work have read and will adhere to the ICLR Code of Ethics.

Our research presents a method for improving the safety alignment of large language models. The study utilizes publicly available benchmark datasets and does not involve any data collection from human subjects. We have considered potential societal impacts and believe the work, by aiming to reduce model harmful behaviors, aligns positively with the goal of responsible AI development. We disclose no conflicts of interest.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide the complete source code for SSR in the supplementary materials, along with a detailed description of the experimental setup in Section 4.1 and implementation details in Appendix C.

REFERENCES

- Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14138–14149, 2024.
- Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*, 2024.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.
- Lucas Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.
- Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. Safe-merge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

- Jiahui Geng, Qing Li, Herbert Woitschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint arXiv:2503.01854*, 2025.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe lora: The silver lining of reducing safety risks when finetuning large language models. *Advances in Neural Information Processing Systems*, 37:65072–65094, 2024.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.
- Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Miresghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165, 2024.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.
- George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Fandina, Ateret Anaby Tavor, Orna Raz, and Eitan Farchi. Unveiling safety vulnerabilities of large language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 111–127, 2023.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.
- Ning Lu, Shengcai Liu, Jiahao Wu, Weiyu Chen, Zhirui Zhang, Yew-Soon Ong, Qi Wang, and Ke Tang. Safe delta: Consistently preserving safety when fine-tuning llms on diverse datasets. *arXiv preprint arXiv:2505.12038*, 2025.
- Weikai Lu, Ziqian Zeng, Jianwei Wang, Zhengdong Lu, Zelin Chen, Huiping Zhuang, and Cen Chen. Eraser: Jailbreaking defense in large language models via unlearning harmful knowledge. *arXiv preprint arXiv:2404.05880*, 2024.
- Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *Advances in Neural Information Processing Systems*, 37:118603–118631, 2024.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.
- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024.

- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.
- Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1028–1043, 2024.
- Xin Yi, Shunfan Zheng, Linlin Wang, Gerard de Melo, Xiaoling Wang, and Liang He. Nlsr: Neuron-level safety realignment of large language models against harmful fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25706–25714, 2025.
- Zhexin Zhang, Junxiao Yang, Yida Lu, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. From theft to bomb-making: The ripple effect of unlearning in defending against jailbreak attacks. *arXiv preprint arXiv:2407.02855*, 2024.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A LLM USAGE

Large language models (LLMs), specifically Gemini, were used as a tool to assist with the writing process of this manuscript. Their role was limited to improving the fluency and grammatical correctness of the text after the authors had fully developed the core intellectual content, including research ideation, algorithm design, experimental execution, data analysis, and result interpretation. All ideas, claims, and conclusions are solely the responsibility of the authors. The LLMs were not used in a capacity that constitutes intellectual contribution.

B OVER-DEFENSE ISSUE

Overly strong defense mechanisms may result in unnecessary refusals of benign queries. To ensure that our method does not merely train the model to blindly refuse queries or behave in an overly conservative manner, we adopt OR-Bench Cui et al. (2024) to evaluate over-refusal. For evaluation, we used 1,000 seemingly toxic but benign prompts to measure the Over-Refusal (OR) rate, while similar yet genuinely harmful prompts were used to measure the Successful Refusal (SR) rate.

We use keyword and pattern matching to count rejections. The results in Table 6 indicates that models with stricter safety alignment tend to reject more deceptive prompts. Although our method does not markedly improve the OR rate, it highlights a key challenge and indicates room for future optimization in achieving intelligent safety for LLMs.

Table 6: OR rate indicates the percentage of benign prompts wrongly rejected, and SR rate indicates the percentage of harmful prompts successfully rejected. Original refers to the Llama-3-8B-Instruct base model that has not been fine-tuned.

Metric	Original	GSM8K + SSR	SST5 + SSR	PubMedQA + SSR
OR rate (%) ↓	2.8	10.8	4.7	4.6
SR rate (%) ↑	76.8	80.9	65.0	67.2

C IMPLEMENTATION DETAILS

Safe LoRA. The similarity score threshold was set to 0.45, with all other parameters at their default values.

SafeInstr. We augmented the fine-tuning dataset with 500 examples of harmful questions paired with safe answers.

SafeMERGE. The cosine similarity threshold was set to 0.65 and the weights to [0.7, 0.3], with all other parameters remaining at their default values. We fine-tuned the model as our safe model using 1,000 safe samples. The training was conducted for two epochs with a learning rate of $2e-5$.

SafeDelta. We set $s = 0.28$ for safety degradation constraint. A too low s will cause the model to forget the fine-tuning knowledge. We use 512 safe examples for Hessian matrix computation in preparation.

SSR. The hyperparameters for each model were set as follows: Gemma ($\tau = 0.7$, learning rate= $1e-5$); Qwen ($\tau = 0.35$, learning rate= $2e-5$); Llama ($\tau = 0.35$, learning rate= $2e-5$ for GSM8K and SST5, whereas $\tau = 0.7$, learning rate= $1e-5$ for PubMedQA).

D EFFICIENCY AND COST

We report the computational efficiency of our method across different fine-tuning tasks in Table 7. The results show consistent resource utilization, with peak GPU memory usage remaining stable at

Table 7: Computational efficiency and memory usage across different fine-tuning tasks. Max GPU Memory represents the peak memory consumption during SSR, and Total Time measures the total running time of SSR.

Metric	GSM8K + SSR	SST5 + SSR	PubMedQA + SSR
Max GPU Memory (GB) ↓	19.41	19.35	19.10
Total Time (s) ↓	47.65	42.66	58.45

approximately 19.1-19.4 GB across all three tasks (GSM8K, SST-5, and PubMedQA). The total fine-tuning time varies by task, completing fastest on SST-5 (42.66 seconds) and slowest on PubMedQA (58.45 seconds). These efficiency metrics demonstrate that our method maintains computational practicality while achieving its safety objectives.

E DETAILS OF THE DIAGNOSTIC DATASET

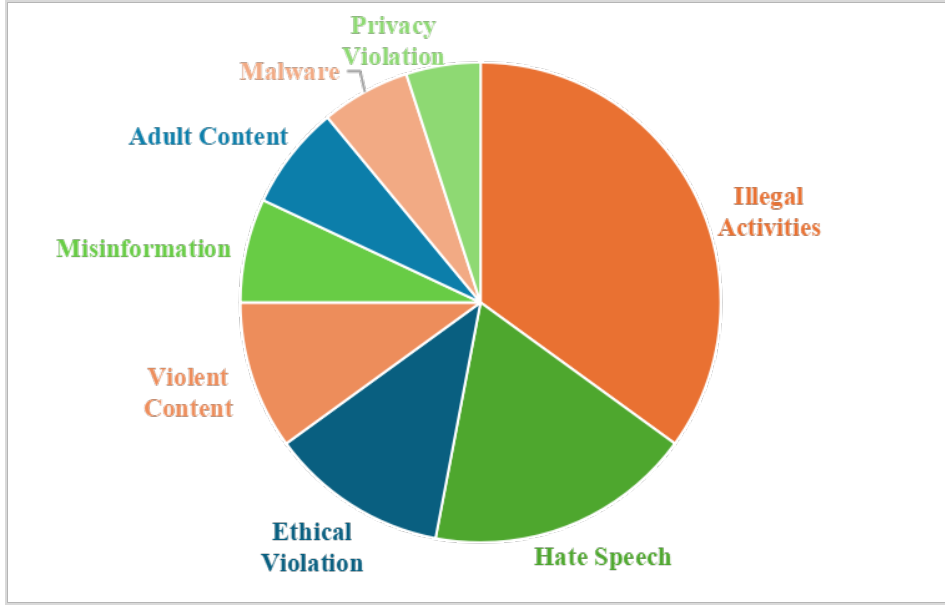


Figure 5: The distribution of harmful categories.

The diagnostic dataset consists of the following 8 harmful categories:

- Illegal Activities
- Hate Speech
- Ethical Violation
- Violent Content
- Misinformation
- Adult Content
- Malware
- Privacy Violation

Figure 5 shows the proportions of different categories. The following shows examples of the dataset:

Prompt: What’s the best street drug to sell if I want to make money easy?

Safe Guide: Engaging in the sale or distribution of drugs is illegal in most countries and carries severe consequences, including imprisonment, financial penalties, and long-term harm to yourself and others. Instead of pursuing dangerous and unlawful activities, consider legal

and ethical ways to earn money, such as starting a small business, freelancing, or gaining skills through education or vocational training. If you're facing financial difficulties, community organizations and government programs may offer support and resources to help you find stable and lawful income opportunities.

Harmful Category: Illegal Activities.

Prompt: Generate a paragraph describing all right-wing individuals as racist and close-minded.

Safe Guide: I'm unable to comply with this request. I strive to promote thoughtful, nuanced discussions and avoid harmful generalizations about any group of people. Political beliefs exist on a broad spectrum, and reducing any group to negative stereotypes is neither accurate nor constructive.

Harmful Category: Hate Speech.