SAFESCHATTEN: SCHATTEN-NORM-CONSTRAINED STEEPEST DESCENT IN ACTIVATION SPACE FOR SAFE LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning can improve an aligned language model's task performance while weakening its safeguards. When safety and task learning depend on overlapping representation directions, protecting safety requires controlling how those directions change during adaptation. We introduce SAFESCHATTEN, a fine-tuning framework that budgets activation drift on trusted safety inputs and chooses updates that maximize local task progress within that budget. Frobenius and spectral constraints provide two ways to distribute the allowance across singular directions, yielding closed-form updates for the linear-layer problem. Across three downstream tasks and five aligned models, SAFESCHATTEN substantially reduces harmfulness while retaining useful adaptation. On SST2, it reduces harmfulness from 51.30% under standard fine-tuning to 11.20%, with accuracy changing from 95.76% to 93.46%. Budget sweeps and update ablations connect these gains to the geometry of activation drift, while a one-step analysis motivates adaptive norm selection. These findings establish activation geometry as a useful design choice for safety-preserving fine-tuning.

1 INTRODUCTION

A model can learn a new task and, in the process, lose some of its ability to refuse harmful requests. Fine-tuning on a small adversarial dataset can weaken an aligned model's safeguards, but the problem also arises during benign adaptation (Qi et al., 2024; Chen et al., 2026). Subsequent work identifies benign examples with unusually damaging fine-tuning effects and constructs attacks using short, benign training trajectories (He et al., 2024; Xie et al., 2025). For a fine-tuning-as-a-service (FTaaS) provider, this creates a difficult tension: customization and adaptation require the model to change, while safety preservation requires some of its behavior to remain stable. The content of a training example alone need not reveal which changes it will induce. In a specific formal setting, Davies et al. (2025) establish limitations of defenses that assess training and inference examples pointwise. Together, these observations motivate looking beyond individual examples to the effect of the training update itself.

Existing defenses provide several ways to intervene. Some strengthen the aligned model against future modification (Huang et al., 2025b; Tamirisa et al., 2025); others incorporate safety supervision during adaptation (Bianchi et al., 2024; Huang et al., 2024) or attenuate harmful gradient influence (Nguyen et al., 2026). A complementary approach constrains updates using safety-related subspaces (Hsu et al., 2024; Zhang et al., 2026). This geometric perspective is appealing because it acts directly on how the model changes. Yet safety and task learning can depend on overlapping directions (Ponkshe et al., 2026): excluding an update from a protected subspace can also exclude useful adaptation. This raises a more flexible design question: how much change should be permitted in safety-relevant representations, and how should that allowance be distributed?

Our hypothesis is that controlling activation drift on trusted safety inputs can retain room for task learning while limiting safety degradation. An aligned-reference activation-patching experiment motivates this focus. Across five models, replacing the fine-tuned model's final-block outputs with the aligned reference's activations reduces mean harmfulness from 48.64% to 7.68%, close to the reference's 7.64%. This further motivates us to preserve safety-relevant computations during training. For fixed inputs $\boldsymbol{X}$ to a linear layer, a weight update changes its outputs by exactly $\Delta \boldsymbol{W}\boldsymbol{X}$, providing a concrete quantity on which to impose a budget.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

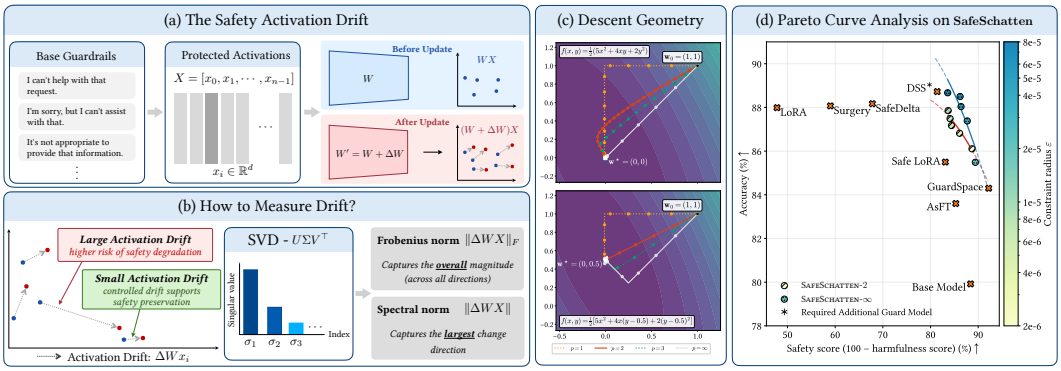


Figure 1: SAFESCHATTEN Overview. (a) On safety inputs $\boldsymbol{X}$, a weight update induces drift $\Delta \boldsymbol{W}\boldsymbol{X}$. (b) Frobenius and spectral norm measure different aspects of its singular-value spectrum. (c) Geometry-dependent descent: two quadratic examples illustrate how the geometry changes the descent trajectory. (d) Pareto Curve Analysis on Llama-3.1-8B-Instruct demonstrates that SAFESCHATTEN reaches better Pareto optimality than baselines on task adaptation (AG News)

We introduce SAFESCHATTEN, which chooses updates that maximize first-order task progress within a budget (ε) on this activation drift. The budget specifies how much change is permitted; its norm specifies how change is allocated across singular directions. A Frobenius norm measures total squared drift energy, whereas a spectral norm measures the largest singular value. Figure 1 illustrates why the same objective can favor different steps under different geometries. Our second hypothesis is that the allocation of change across these directions affects the safety-utility tradeoff. This motivates treating the norm as part of the optimizer design. These two members of the Schatten-norm family yield closed-form updates that combine the safety-activation covariance with different weightings of the transformed gradient's singular values. This makes the geometry of protected activation space an explicit part of the optimizer.

Experiments spanning three downstream tasks and five aligned models demonstrate strong safety-utility performance. On SST2, the spectral variant reduces harmfulness from 51.3% under standard fine-tuning to 11.2%, while retaining 93.46% accuracy versus 95.76%. On GSM8K, it achieves 82.11% accuracy with 16.0% harmfulness, compared with 83.21% accuracy and 49.3% harmfulness for standard fine-tuning. Figure 1(d) further shows that varying the drift budget enables SAFESCHATTEN to reach Pareto-optimal safety-utility operating points among the evaluated methods on AG News with Llama-3.1-8B-Instruct, without requiring an auxiliary guard model.

In addition to the standard SAFESCHATTEN variants based on the Frobenius and spectral norms, we find that adaptively selecting the appropriate norm over the course of training can further improve performance. Together, these results develop a practical principle for safer adaptation: optimize task progress within a deliberately chosen budget on safety-relevant representational change.

The paper thus makes three contributions:

- **A geometric formulation of safety-preserving adaptation.** We formulate fine-tuning as task optimization within a budget on activation drift over trusted safety inputs. The norm determines how this budget is distributed across singular directions, allowing controlled adaptation in directions relevant to both safety and utility.

- **Closed-form optimizers for protected activation space.** We derive Frobenius- and spectral-constrained steepest-descent updates, linking the safety-activation covariance and gradient spectrum to the optimizer's update direction. We also conduct a one-step analysis that motivates adaptive selection between the two geometries.

- **Strong safety-utility performance.** Across three downstream tasks and five aligned models, SAFESCHATTEN preserves substantial task gains while reducing harmfulness. On SST2, it reduces harmfulness by 40.1% relative to standard fine-tuning for a 2.3% accuracy cost; matrix-sign ablations and budget sweeps connect performance to the update geometry.

2 SAFESCHATTEN: FROM ACTIVATION BUDGETS TO DESCENT DIRECTIONS

Notation. For a vector $\boldsymbol{v}$, we denote its l_2 norm as $\|\boldsymbol{v}\|_2$. For a matrix $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, we denote its spectral norm as $\|\boldsymbol{A}\|$, Frobenius norm as $\|\boldsymbol{A}\|_{\mathrm{F}}$, nuclear norm as $\|\boldsymbol{A}\|_*$, where $\|\boldsymbol{A}\|_* = \sum_{i=1}^{r}\sigma_i$. Here, $\{\sigma_i\}_{i=1}^{r}$ are singular values. For $p > 1$, we denote Schatten$-p$ norm as $\|\boldsymbol{A}\|_{\mathcal{S}_p}$, where $\|\boldsymbol{A}\|_{\mathcal{S}_p} = (\sum_{i=1}^{r}\sigma_i^p)^{\frac{1}{p}}$. We use $\langle \boldsymbol{A}, \boldsymbol{B}\rangle$ to denote the inner product between $\boldsymbol{A}$ and $\boldsymbol{B}$, i.e., $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{Tr}\left(\boldsymbol{A}^\top \boldsymbol{B}\right)$.

2.1 WHAT SHOULD A SAFE UPDATE PRESERVE?

Parameter movement alone does not tell us how strongly an update changes the model on safety-relevant inputs. For a linear layer, the same update $\Delta \boldsymbol{W}$ can have a large effect on one input direction and little effect on another. Our hypothesis is that measuring the induced activation change provides a useful basis for protecting safety behavior.

Motivating Causal Intervention. We first test whether restoring an activation changed by harmful fine-tuning can restore safety behavior. We fine-tune five aligned models for three epochs on 100 harmful prompt-compliance pairs (Qi et al., 2024), then evaluate on 1,000 harmful prompts from BeaverTails (Ji et al., 2023) on three conditions: the aligned model (*Base*), the adapted model (*Fine-tuned*), and the adapted model with aligned-reference activation patching (ARAP). For ARAP, we replace the fine-tuned model's final decoder-block outputs with cached outputs from the aligned model (*Base*) at matched token positions.

To quantify the *safety activation drift* (SAD) caused by fine-tuning, we compare the base and fine-tuned activations on the same base-generated token trajectory. Let H_i^0 and H_i be the resulting base and fine-tuned final-block activation matrices for prompt i, we define the related Frobenius drift $\|\mathrm{SAD}\|_F$ and spectral drift $\|\mathrm{SAD}\|$:

$$\|\mathrm{SAD}\|_F = \frac{1}{m}\sum_{i=1}^{m}\frac{\|H_i - H_i^0\|_F}{\|H_i^0\|_F}, \qquad \|\mathrm{SAD}\| = \frac{1}{m}\sum_{i=1}^{m}\frac{\|H_i - H_i^0\|}{\|H_i^0\|}. \tag{1}$$

We also report the mean cosine similarity between their flattened matrices, $\frac{1}{m}\sum_i \langle H_i, H_i^0\rangle/(\|H_i\|_F\|H_i^0\|_F)$. As shown in Table 1, harmful fine-tuning increases mean harmfulness from 7.64% to 48.64%, whereas ARAP reduces it to 7.68%, recovering nearly all of the lost safety behavior. Fine-tuning also produces substantial activation drift under both measures, with mean relative Frobenius and spectral drifts of 0.6652 and 0.9654, respectively. These drift measures capture changes that can be obscured by high cosine similarity. For example, for Qwen-2-7B-Instruct, the cosine similarity remains 0.9165 despite a relative Frobenius drift of 0.4486 and a relative spectral drift of 0.5929.

This activation intervention experiment demonstrates a simple yet effective way to preserve models' safety. This further motivates the testable design hypothesis pursued below: constrained activation changes during learning may preserve safety while still permitting task adaptation.

Table 1: Aligned-reference activation patching (ARAP) recovers near-reference harmfulness

Model	Harmfulness ↓			FT activation vs. Base		
	Base	FT	FT (ARAP)	Cosine ↑	$\|\mathrm{SAD}\|_F$ ↓	$\|\mathrm{SAD}\|$ ↓
Llama-2-7B-Chat	0.90	28.80	1.00	0.8938	0.5095	0.9287
Llama-3.1-8B-Instruct	8.70	54.10	8.50	0.6946	1.0189	1.4897
Llama-3.2-3B-Instruct	16.60	46.90	16.40	0.8363	0.6581	0.9173
Qwen-2-7B-Instruct	6.40	48.30	6.40	0.9165	0.4486	0.5929
Gemma-2-9B-IT	5.60	65.10	6.10	0.8812	0.6907	0.8985
Mean	7.64	48.64	7.68	0.8444	0.6652	0.9654

Problem Formulation. Let $\mathcal{D}$ be the downstream fine-tuning dataset, and let $\mathcal{H}$ denote the collection of input activation matrices obtained from a trusted safety alignment dataset. For a weight matrix $\boldsymbol{W} \in \mathbb{R}^{d_{\mathrm{out}}\times d_{\mathrm{in}}}$ and input safety activation $\boldsymbol{X} \in \mathbb{R}^{d_{\mathrm{in}}\times N}$, where N is the number of tokens, write $\mathcal{A}(\boldsymbol{X}, \boldsymbol{W})$ for the corresponding activation matrix. Motivated by prior work on activation-space

safety preservation (Zhang et al., 2026), a natural starting point is a Frobenius drift budget:

$$\min_{\Delta \boldsymbol{W}} \mathcal{L}_{\text{task}}(\boldsymbol{W} + \Delta \boldsymbol{W}; \mathcal{D}), \quad \text{s.t.} \quad \|\mathcal{A}(\boldsymbol{X}, \boldsymbol{W} + \Delta \boldsymbol{W}) - \mathcal{A}(\boldsymbol{X}, \boldsymbol{W})\|_F \leq \eta\varepsilon, \quad \forall \boldsymbol{X} \in \mathcal{H}. \tag{2}$$

Here $\eta > 0$ is the step scale and $\varepsilon \geq 0$ controls the drift allowance per unit step. This formulation gives adaptation an explicit budget on the chosen activations. Whether that budget also preserves refusal behavior is a separate empirical question.

2.2 How should the drift budget be measured?

In this work, we study optimization in the activation space relevant to model safety from a broader geometric perspective. Rather than restricting ourselves to the conventional Euclidean geometry, we consider a more general family of geometries induced by different Schatten norm choices in (3), note that $p = 2$ returns to problem (2).

$$\min_{\Delta \boldsymbol{W}} \mathcal{L}_{\text{task}}(\boldsymbol{W} + \Delta \boldsymbol{W}; \mathcal{D}), \quad \text{s.t.} \quad \|\mathcal{A}(\boldsymbol{X}, \boldsymbol{W} + \Delta \boldsymbol{W}) - \mathcal{A}(\boldsymbol{X}, \boldsymbol{W})\|_{\mathcal{S}_p} \leq \eta\varepsilon, \quad \forall \boldsymbol{X} \in \mathcal{H}, \tag{3}$$

where we assume that activation looks like $\mathcal{A}(\boldsymbol{X}, \boldsymbol{W}) = \sigma(\boldsymbol{W}\boldsymbol{X})$ and $\sigma(\cdot)$ is the activation function. This perspective allows us not only to examine optimization under the Euclidean geometry, but also to explore alternative geometric structures that may be equally well suited (or even potentially better suited) to optimizing safety-related representations compared to Euclidean geometry.

One may naturally ask why we focus on Schatten-norm constraints rather than element-wise norm constraints. The main motivation is that, as we will show in Lemma 1, steepest descent under a Schatten-norm constraint induces a spectral update in the eigenspace, rather than acting independently on individual entries. this spectral structure allows the update to adapt to the geometry of the parameter space and inducing an implicit preconditioning effect (Deng et al., 2026; Wang et al., 2026). While entry-wise transformations may also introduce implicit preconditioning, the spectral formulation acts jointly on the parameter matrix rather than treating individual entries independently, thereby explicitly accounting for interdependencies among parameter directions. Such matrix-level coupling can reshape the update along coordinated directions in parameter space, which may be better suited to capturing the geometry relevant to both downstream optimization and safety preservation. In Section 3.4.1, we further compare Schatten-norm steepest descent with its element-wise counterpart and study how their induced update geometries differ.

2.3 Why does geometry change the descent direction?

Before we start, we first consider the following simplified constrained minimization problem:

$$\min_{\Delta \boldsymbol{W}} f(\boldsymbol{W} + \Delta \boldsymbol{W}), \quad \text{s.t.} \quad \|\Delta \boldsymbol{W}\|_{\mathcal{S}_p} \leq \eta\varepsilon. \tag{4}$$

When we consider the first-order approximation for the objective in (4), we have

$$f(\boldsymbol{W} + \Delta \boldsymbol{W}) \approx f(\boldsymbol{W}) + \text{Tr}(\nabla f(\boldsymbol{W})^\top \Delta \boldsymbol{W}) \tag{5}$$

It is straightforward to derive that optimizing this approximation is equivalent to the following objective:

$$\max_{\Delta \boldsymbol{W}} -\text{Tr}(\nabla f(\boldsymbol{W})^\top \Delta \boldsymbol{W}), \quad \text{s.t.} \quad \|\Delta \boldsymbol{W}\|_{\mathcal{S}_p} \leq \eta\varepsilon. \tag{6}$$

To better build intuition, here we make use of this simple optimization problem to illustrate how steepest descent under non-Euclidean geometries can lead to qualitatively different update directions, and may potentially offer advantages over conventional Euclidean steepest descent. Before we start, we need to consider the following simplified constrained minimization problem in (4). We have Lemma 1 for obtaining the update expression. We put in the proof of Lemma 1 in Appendix A.2.

Lemma 1. *Let $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the SVD of the negative gradient $\boldsymbol{G} = -\nabla f(\boldsymbol{W})$, where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_r)$. Let p be the conjugate exponent of q, satisfying $1/p + 1/q = 1$. The explicit solution $\Delta \boldsymbol{W}^*$ to the optimization problem 6 is given by:*

$$\Delta \boldsymbol{W}^* = -\eta\varepsilon \frac{1}{\|\boldsymbol{G}\|_{\mathcal{S}_q}^{q/p}} \boldsymbol{U}\boldsymbol{\Sigma}^{q-1}\boldsymbol{V}^\top, \tag{7}$$

where $\boldsymbol{\Sigma}^{q-1} = \text{diag}(\sigma_1^{q-1}, \ldots, \sigma_r^{q-1})$.

Based on (4), if we consider a one-dimension case where $f(\boldsymbol{w}) = f(x, y, z) = \frac{1}{2}(9x^2 + 4y^2 + z^2)$ with $(x_0, y_0, z_0) = (1, 1, 1)$. Viewing this toy problem as a three-dimensional diagonal setting, we have $\boldsymbol{G} = \mathrm{diag}(9, 4, 1)$, so its singular values are $(\sigma_1, \sigma_2, \sigma_3) = (9, 4, 1)$; while the ordinary gradient update is proportional to $-(9, 4, 1)$, Lemma 1 gives a Schatten$-p$ steepest-descent update proportional to $-(9^{q-1}, 4^{q-1}, 1)$, which points directly from $(1, 1, 1)$ toward the minimizer only when it is proportional to $-(1, 1, 1)$. This means $p = \infty$ yields the fastest convergence for this problem under steepest-descent updates, this example suggests that the Schatten-∞ geometry may be better matched to the optimization landscape at the given initialization. We also give a similar visualization in Figure 1(c). Therefore, in the context of optimizing safety-related activation spaces, we should not rule out the possibility that a Schatten geometry other than the conventional Schatten-2 (Frobenius) geometry may provide a more suitable choice. Geometrically, different Schatten norms induce different feasible update geometries and therefore better or worse match the local optimization landscape.

2.4 From geometric intuition to activation-constrained updates

The above example explains how a norm reshapes a step. When we employ the first-order approximation for our optimization objective

$$\mathcal{L}_{\text{task}}(\boldsymbol{W} + \Delta\boldsymbol{W}; \mathcal{D}) \approx \mathcal{L}_{\text{task}}(\boldsymbol{W}; \mathcal{D}) + \langle \nabla\mathcal{L}_{\text{task}}(\boldsymbol{W}; \mathcal{D}), \Delta\boldsymbol{W} \rangle \tag{8}$$

It is straightforward to derive that optimizing this approximation is equivalent to the following objective:

$$\min_{\Delta\boldsymbol{W}} \langle \nabla\mathcal{L}_{\text{task}}(\boldsymbol{W}; \mathcal{D}), \Delta\boldsymbol{W} \rangle, \quad \text{s.t.} \quad \|\mathcal{A}(\boldsymbol{X}, \boldsymbol{W} + \Delta\boldsymbol{W}) - \mathcal{A}(\boldsymbol{X}, \boldsymbol{W})\|_{\mathcal{S}_p} \le \eta\varepsilon, \quad \forall \boldsymbol{X} \in \mathcal{H}. \tag{9}$$

Now we denote $\boldsymbol{G}_{\text{task}} = \nabla\mathcal{L}_{\text{task}}(\boldsymbol{W}; \mathcal{D})$ for layer $\boldsymbol{W}$ and $\Delta\boldsymbol{W} = -\eta\boldsymbol{\Phi}$, we have

$$\max_{\boldsymbol{\Phi}} \langle \boldsymbol{G}_{\text{task}}, \boldsymbol{\Phi} \rangle, \quad \text{s.t.} \quad \|\mathcal{A}(\boldsymbol{X}, \boldsymbol{W} - \eta\boldsymbol{\Phi}) - \mathcal{A}(\boldsymbol{X}, \boldsymbol{W})\|_{\mathcal{S}_p} \le \eta\varepsilon, \quad \forall \boldsymbol{X} \in \mathcal{H}. \tag{10}$$

Where $\boldsymbol{W} \in \mathbb{R}^{d_{out} \times d_{in}}$, $\boldsymbol{X} \in \mathbb{R}^{d_{in} \times N}$, N is the number of tokens.

Assumption 1. *We assume that the activation function satisfies $m\boldsymbol{I}_{d_{in} \times N} \le \sigma'(\cdot) \le M\boldsymbol{I}_{d_{in} \times N}$, where $0 < m \le M < \infty$ are constants.*

Theorem 1. *Let $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top$ defined as the safety-activation covariance. Assume that $\boldsymbol{K} \succ \boldsymbol{0}$. Then, for $p \in \{2, \infty\}$, under Assumption 1, the layer-wise steepest descent directions for optimization problem (10) at each step are given by*

$$\boldsymbol{\Phi}_p^\star = \begin{cases} \varepsilon \dfrac{\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}}{\left\|\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}\right\|_F}\boldsymbol{K}^{-1/2}, & p = 2, \\ \varepsilon\ \mathrm{msign}\left(\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}\right)\boldsymbol{K}^{-1/2}, & p = \infty. \end{cases} \tag{11}$$

for an SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, we define $\mathrm{msign}(\boldsymbol{A}) := \boldsymbol{U}\boldsymbol{V}^\top$.

In this work, we mainly consider $p \in \{2, \infty\}$,. By Theorem 1, we obtain the steepest descent update under different Schatten norm constraint, this leads to SafeSchatten in Algorithm 1. We put the proof of Theorem 1 in Appendix A.2.

Algorithm 1 SafeSchatten$-p$

1: **Input:** Initial weights $\mathbf{W}_0 \in \mathbb{R}^{d_{out} \times d_{in}}$, Downstream task loss function $\mathcal{L}_{\text{task}}$, Safety alignment data $\boldsymbol{X} \in \mathbb{R}^{d_{in} \times N}$, $p \in \{2, \infty\}$, learning rate η, weight decay λ.
2: $\boldsymbol{K} \leftarrow \boldsymbol{X}\boldsymbol{X}^\top$
3: **for** $t = 1, 2, \ldots$ **do**
4: $\quad \mathbf{G}_t \leftarrow \nabla_{\mathbf{W}}\mathcal{L}_{\text{task}}(\mathbf{W}_t)$
5: $\quad \mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \eta\lambda\mathbf{W}_t - \eta\boldsymbol{\Phi}_p^\star$ ▷ We can select $\boldsymbol{\Phi}_p^\star$ from Theorem 1
6: **end for**

3 Experiments

We evaluate whether SafeSchatten preserves safety while learning a downstream task, how this tradeoff changes with the amount of fine-tuning data and its harmful fraction, and whether the spectral update benefits from operating on matrix singular directions. We denote SafeSchatten-2 for using the Frobenius norm as the constraint and SafeSchatten-∞ for spectral norm.

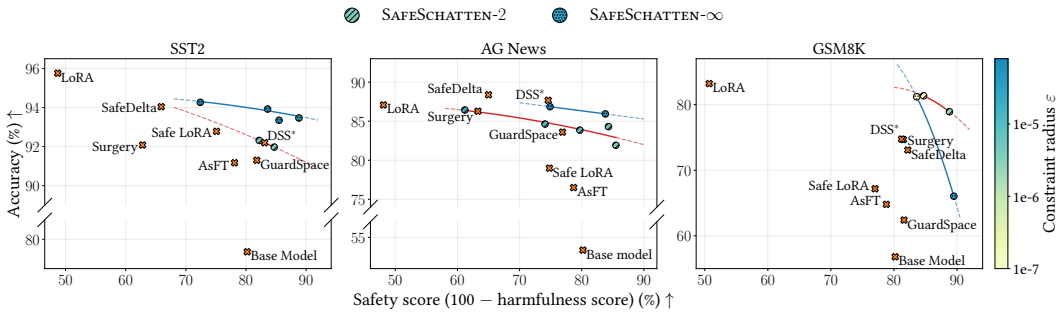


Figure 2: Pareto Curve Analysis on SafeSchatten. Llama-3.2-3B across three SST2, AG News, and GSM8K. Circles show budget sweeps; crosses show reference settings. Safety is $100 - \mathrm{Harm}$; upper right is better. DSS* uses an auxiliary guard model

3.1 Experimental Setup

Datasets and Models. We consider three fine-tuning datasets: SST2 (Socher et al., 2013), AG News (Zhang et al., 2015), and GSM8K (Cobbe et al., 2021). Within a total of n samples, we mix ρ (percentage) of unsafe data from BeaverTails (Ji et al., 2023) with the fine-tuning task's *benign* training data. In our experiment, the default setting is $\rho = 0.10$ and $n = 1,000$ unless otherwise specified. We use five aligned models: Llama-3.2-3B-Instruct (Grattafiori et al., 2024), Llama-2-7B-Chat (Touvron et al., 2023), Qwen2-7B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Gemma-2-9B-Instruct (GemmaTeam et al., 2024) in our experiments.

Baselines. We use seven baselines for comparison. LoRA (Hu et al., 2022) is the vanilla supervised fine-tuning. Safe LoRA (Hsu et al., 2024) and AsFT (Yang et al., 2026) are two fine-tuning stage defenses based on LoRA adapters. Surgery (Liu et al., 2026), GuardSpace (Zhang et al., 2026), and DSS (Peng et al., 2025) are three dense fine-tuning-stage defenses that update more model parameters (i.e., high-rank or full-parameter); it is also worth noting that DSS utilizes an additional guard model (Llama-Guard-3-8B) in defense. We also include a post-hoc defense, SafeDelta (Lu et al., 2025), in the experiment.

Training Details and Metrics. Unless otherwise specified, we follow Hsu et al. (2024) and Yang et al. (2026) to utilize LoRA (Hu et al., 2022) for parameter-efficient training. The rank of the adaptor is set to 8, and the LoRA's alpha is the same as the rank. For fine-tuning tasks, we fine-tune SST2, AGNEWS, and GSM8K for ten epochs. See Appendix A.3 for more training details. For evaluation, we report task accuracy (*Acc.*, %) and harmfulness score (*Harm*, %) on 1,000 harmful prompts from the BeaverTails dataset (Ji et al., 2023), in which we employ the moderation model to classify the model output as harmful or not. Therefore, *Harm* is the percentage of flagged responses. To measure safety activation change, we compare the final decoder-block outputs of the adapted and aligned models on a small safety dataset. We then report mean relative Frobenius and spectral drift: $\|\mathrm{SAD}\|_F$ and $\|\mathrm{SAD}\|$, as defined in Eq 1.

3.2 Safety and Downstream Utility

Pareto Optimal. Figure 2 shows how varying the activation-drift budget ε allows SafeSchatten to balance task accuracy and safety. Across SST2, AG News, and GSM8K, the sweeps reveal operating points that improve both metrics over several reported defense settings. We define the safety score as $100 - \mathrm{Harm}$, so movement toward the upper right improves both objectives. The curves trace SafeSchatten across budgets. It is worth noting that SafeSchatten is Pareto optimal when only the base model is utilized to preserve model safety during fine-tuning, where DSS* requires additional guard models is excluded.

Substantial safety gains with strong task performance. Table 2 quantifies the gains at representative configurations. Compared with standard fine-tuning, the SafeSchatten$-\infty$ reduces SST2 harmfulness from 51.3% to 11.2%, while retaining 93.46% accuracy versus 95.76%. This is a 40.1-percentage-point reduction in harmfulness for a 2.3-point accuracy cost. On GSM8K, it reduces harmfulness from 49.3% to 16.0% with only a 1.1-point accuracy cost (83.21% to 82.11%); both SafeSchatten-2 and SafeSchatten$-\infty$ also achieve higher accuracy and lower harmfulness

Table 2: Safety and task performance after fine-tuning on SST2, AG News, and GSM8K

	SST2				AG News				GSM8K			
	Harm (↓)	Acc. (↑)	$\|\text{SAD}\|_F$	$\|\text{SAD}\|$	Harm (↓)	Acc. (↑)	$\|\text{SAD}\|_F$	$\|\text{SAD}\|$	Harm (↓)	Acc. (↑)	$\|\text{SAD}\|_F$	$\|\text{SAD}\|$
Base Model	19.80	79.36	-	-	19.80	53.37	-	-	19.80	56.80	-	-
LoRA	51.30	95.76	0.4670	0.3483	51.90	87.08	0.4278	0.2917	49.30	83.21	0.3876	0.2443
SafeDelta	34.10	94.04	0.4438	0.3195	35.00	88.38	0.4418	0.3140	17.80	73.08	0.3186	0.1905
DSS	16.90	92.20	1.7508	2.2460	25.40	87.68	1.6528	2.0783	18.50	74.68	1.0636	1.2256
Surgery	37.20	92.08	0.7305	0.4415	36.70	86.28	0.7480	0.4737	18.80	74.75	0.0197	0.0077
Safe LoRA	24.90	92.78	0.1304	0.0876	25.20	78.99	0.1320	0.0791	23.00	67.17	0.1728	0.0936
AsFT	21.90	91.17	0.1225	0.0652	21.30	76.50	0.1125	0.0552	21.20	64.80	0.1368	0.0611
GuardSpace	18.20	91.30	0.4465	0.3060	23.10	83.60	0.4421	0.2948	18.40	62.40	0.3478	0.2362
SAFESCHATTEN−2	13.60	92.54	0.1000	0.0614	15.70	84.30	0.2811	0.2024	15.30	81.35	0.0187	0.0067
SAFESCHATTEN−∞	11.20	93.46	0.1359	0.0921	16.20	85.93	0.2704	0.1982	16.00	82.11	0.0181	0.0062

Table 3: Safety and task performance after fine-tuning five models on AG News

	Llama-3.2-3B-Instruct		**Llama-2-7B-Chat**		**Qwen2-7B-Instruct**		**Llama-3.1-8B-Instruct**		**Gemma2-9B-Instruct**	
	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)
Base Model	19.80	53.37	1.20	68.79	7.80	78.45	11.60	79.92	2.50	83.40
LoRA	51.90	87.08	51.40	88.17	51.40	87.43	52.20	87.99	56.20	87.92
SafeDelta	35.00	88.38	10.40	89.34	21.70	87.64	32.20	88.17	15.10	87.59
DSS	25.40	87.68	12.20	88.04	25.40	88.07	18.60	88.72	11.70	88.61
Surgery	36.70	86.28	16.60	88.38	61.10	86.17	41.00	88.07	48.40	87.92
Safe LoRA	25.20	78.99	19.30	76.47	7.40	85.53	16.90	85.50	8.50	86.70
AsFT	21.30	76.50	12.60	86.40	8.00	85.60	14.70	83.60	9.00	89.80
GuardSpace	23.10	83.60	3.60	89.10	24.70	85.50	31.00	89.50	22.90	89.40
SAFESCHATTEN−2	15.70	84.30	1.20	86.74	7.50	87.34	16.30	87.86	5.20	88.18
SAFESCHATTEN−∞	16.20	85.93	1.10	81.95	7.70	86.99	16.40	88.67	5.60	85.89

than every evaluated safety defense. Across all three tasks, both SAFESCHATTEN variants achieve competitive task accuracy while keeping harmfulness below the original (Base Model) 19.8% level.

Safety gains extend across model families. Table 3 evaluates AG News adaptation across five aligned models. Both SAFESCHATTEN−2 and SAFESCHATTEN−∞ reduce harmfulness relative to standard fine-tuning on every model while improving task accuracy to a competitive level. On average, SAFESCHATTEN-2 reduces harmfulness from 52.62% to 9.18%, while retaining 86.88% accuracy compared with LoRA's 87.72%; likewise, SAFESCHATTEN-∞ reduces harmfulness to 9.4%, while retaining 85.89% accuracy. These results show that the safety gains extend beyond a single model and can be achieved with a small average accuracy cost.

Activation drift and geometry explain the design choices. Both SAFESCHATTEN−2 and SAFESCHATTEN−∞ reduce relative Frobenius and spectral activation drift compared with LoRA on all three tasks (Table 2). On GSM8K, SAFESCHATTEN−∞ reduces Frobenius drift from 0.3876 to 0.0181 and spectral drift from 0.2443 to 0.0062, alongside its strong safety and accuracy results. Together, these results support our design: constraining changes to safety-relevant activations helps preserve safe behavior while retaining task gains. However, DSS appears to be an exception, showing larger activation drift while maintaining lower harmfulness. This may reflect its distinct mechanism, which uses an auxiliary guard model for token-level loss shaping (Peng et al., 2025) rather than directly controlling activation drift. The choice of geometry also matters: at the reported configurations, SAFESCHATTEN−∞ achieves lower harmfulness and higher accuracy than SAFESCHATTEN−2 on SST2, whereas SAFESCHATTEN−2 does so on AG News. These task-dependent outcomes motivate treating geometry as an optimizer design choice and investigating adaptive norm selection in Section 3.4.3.

3.3 Sensitivity to Data Size and Harmful Fraction

On AG News with Llama-3.2-3B-Instruct, SAFESCHATTEN−2 and SAFESCHATTEN−∞ remain competitive low harmfulness scores while retaining useful task adaptation in separate sweeps of training-set size ($n = 0.5\text{K} - 2.5\text{K}$ examples) and harmful fraction ($\rho = 0 - 20\%$). Due to the space limit, we provide the experimental setup, full results, and analysis in Appendix A.4.

3.4 Which Update Choices Matter?

3.4.1 Matrix-Sign versus Elementwise-Sign Updates

In this section, we empirically investigate whether steepest descent under a Schatten-norm constraint provides advantages over element-wise updates. Specifically, we take the comparison between SAFESCHATTEN−∞ and SAFESIGNGD. For SAFESIGNGD, we consider the update

Table 5: Performance on incorporating momentum

	SST2		AG News	
	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)
Base Model	19.80	79.36	19.80	53.37
LoRA	51.30	95.76	51.90	87.08
SAFESCHATTEN-2	13.60	92.54	15.70	84.30
SAFESCHATTEN-2 (M)	15.20	92.89	17.80	85.36
SAFESCHATTEN-∞	11.20	93.46	15.90	81.42
SAFESCHATTEN-∞ (M)	25.50	94.04	29.80	84.78

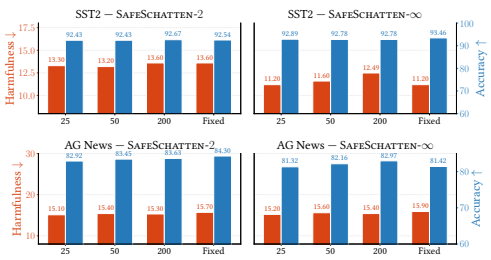


Figure 3: Safety activation update freq.

$\mathbf{\Phi} = \varepsilon \frac{\mathrm{Sign}(\boldsymbol{G}_{\mathrm{task}} \boldsymbol{K}^{-1/2})}{\left\| \boldsymbol{G}_{\mathrm{task}} \boldsymbol{K}^{-1/2} \right\|_F} \boldsymbol{K}^{-1/2}$. which applies the sign operation element-wise, as a representative example. We want to test whether the spectral variant benefits from shaping a matrix's singular directions by replacing its matrix-sign operation with element-wise sign.

As shown in Table 4, SAFESCHATTEN-∞ improves both reported metrics on each task. On SST2, it reduces harmfulness from 18.20% to 11.20% while increasing accuracy from 92.55% to 93.46%. The largest utility gain occurs on GSM8K, where accuracy increases from 75.59% to 82.11% and harmfulness decreases from 17.90% to 16.00%. These results suggest that, under the same optimization budget, the spectral steepest-descent update provides a more effective update geometry than its element-wise counterpart, improving both utility and safety across the evaluated tasks.

Table 4: Comparsion between SAFESCHATTEN$-\infty$ and SAFESIGNGD on different datasets

	SST2		AG News		GSM8K	
	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)
SAFESCHATTEN$-\infty$	11.2	93.46	15.9	81.42	16.0	82.11
SAFESIGNGD	18.2	92.55	16.4	78.82	17.9	75.59

3.4.2 MOMENTUM AND ACTIVATION-CACHE REFRESH

As an additional ablation, we consider incorporating momentum into the downstream fine-tuning gradient $\boldsymbol{G}_{\mathrm{task}}$. Since momentum aggregates gradient information across optimization steps, it may modify the effective update direction and, consequently, the allocation of the activation budget. We evaluate whether our method exhibits similar behavior under this variant. Specifically at each step t, we consider $\boldsymbol{M}^t_{\mathrm{task}} = \beta_1 \boldsymbol{M}^{t-1}_{\mathrm{task}} + (1-\beta_1)\boldsymbol{G}^t_{\mathrm{task}}$, then we replace $\boldsymbol{G}^t_{\mathrm{task}}$ with $\boldsymbol{M}^t_{\mathrm{task}}$ in (11) to get

$$\mathbf{\Phi}^{\star t}_p = \begin{cases} \varepsilon \dfrac{\boldsymbol{M}^t_{\mathrm{task}} \boldsymbol{K}^{-1/2}}{\left\| \boldsymbol{M}^t_{\mathrm{task}} \boldsymbol{K}^{-1/2} \right\|_F} \boldsymbol{K}^{-1/2}, & p = 2, \\ \varepsilon \ \mathrm{msign}\left(\boldsymbol{M}^t_{\mathrm{task}} \boldsymbol{K}^{-1/2} \right) \boldsymbol{K}^{-1/2}, & p = \infty. \end{cases} \tag{12}$$

We evaluate SAFESCHATTEN-2(M) and SAFESCHATTEN$-\infty$(M) with Llama-3.2-3B-Instruct on SST2 and AG News datasets. In Table 5, we can see that, compared with SAFESCHATTEN-2 or SAFESCHATTEN$-\infty$, SAFESCHATTEN-2(M) or SAFESCHATTEN$-\infty$(M) achieves slightly higher downstream-task accuracy while also yielding slightly lower harmfulness scores, reflecting a certain trade-off between the two objectives. We adopt $\beta_1 = 0.9$ for SAFESCHATTEN-2(M) and $\beta_1 = 0.95$ SAFESCHATTEN$-\infty$(M).

Must the safety activations track the evolving model? In our main experiments, for computational efficiency, we keep the input safety activations fixed throughout fine-tuning, using activations computed from the initial safety-aligned base model. Here, we further examine whether periodically refreshing these activations with the current model during training can provide additional benefits. We study it by comparing a fixed cache with refreshes every 25, 50, or 200 steps. Figure 3 shows no significant improvement from more frequent refreshes. Across the plotted settings, harmfulness ranges from 13.2% to 13.6% for SAFESCHATTEN-2 on SST2 and from 15.1% to 15.7% on AG News; the spectral-norm constraint also shows modest, non-monotone changes. The reported results support a fixed safety activation as a practical default for these runs.

3.4.3 HOW TO CHOOSE SCHATTEN$-p$ NORM

Building on SAFESCHATTEN, a natural extension is to ask at step t, when the update for each layer should follow the steepest descent direction induced by a Schatten-2 constraint versus that induced

Table 6: Performance on adaptive norm selection

	SST2		AG News	
	Harm (↓)	Acc. (↑)	Harm (↓)	Acc. (↑)
Base Model	19.80	79.36	19.80	53.37
SFT	54.30	95.18	53.70	89.04
SAFESCHATTEN-2	17.70	93.46	42.00	82.92
SAFESCHATTEN-∞	25.90	94.61	31.00	79.24
SAFESCHATTEN-Adaptive	24.30	94.72	32.20	83.25

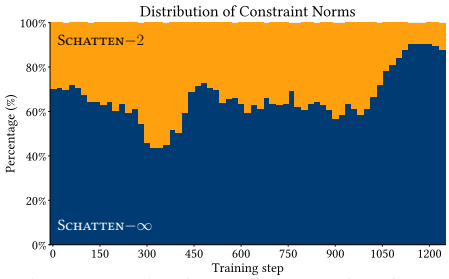


Figure 4: Adaptive norm selection freq.

by a Schatten-∞ constraint. To solve this, we follow Davis & Drusvyatskiy (2025) and Massena et al. (2026) and analyze the problem from a layer-by-layer perspective. Our starting point is to select the $\mathcal{S}_p$ constraint that yields the largest decrease in $\mathcal{L}_{\text{task}}$ after a single step. To make the analysis tractable, we consider a simplified single-layer feature regression loss that serves as a surrogate for $\mathcal{L}_{\text{task}}$:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2N}\|\boldsymbol{W}\mathcal{A} - \boldsymbol{Y}^*\|_F^2, \tag{13}$$

where Y^* denotes the *target* representation corresponding to this layer for downstream finetuning data and $\mathcal{A}$ is the input activation for downstream finetuning data .

If we update $\boldsymbol{W}$ by $\Delta\boldsymbol{W}$, we have the exact expansion: $\mathcal{L}(\boldsymbol{W}+\Delta\boldsymbol{W}) = \mathcal{L}(\boldsymbol{W}) + \langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle + \frac{1}{2N}\|\Delta\boldsymbol{W}\mathcal{A}\|_F^2$, where $\boldsymbol{G} = \nabla\mathcal{L}(\boldsymbol{W})$. The next step is to analyze how the steepest descent updates under the $\mathcal{S}_2$ and $\mathcal{S}_\infty$ norm constraints affect the one-step reduction in the loss.

Theorem 2 (Choosing between $\mathcal{S}_2$ and $\mathcal{S}_\infty$ constraints)**.** *Suppose we consider a simplified surrogate loss in (13) Let* $\boldsymbol{B} := \operatorname{msign}\left(\boldsymbol{G}\boldsymbol{K}^{-1/2}\right)\boldsymbol{K}^{-1/2}$. *For the optimal $\mathcal{S}_2$ and $\mathcal{S}_\infty$ update directions, optimizing over the step size η gives* $\Delta\mathcal{L}^\star_{\mathcal{S}_2} = \frac{N}{2}\frac{\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\|_F^4}{\|\boldsymbol{G}\boldsymbol{K}^{-1}\mathcal{A}\|_F^2}$, $\Delta\mathcal{L}^\star_{\mathcal{S}_\infty} = \frac{N}{2}\frac{\langle \boldsymbol{G},\boldsymbol{B}\rangle^2}{\|\boldsymbol{B}\mathcal{A}\|_F^2}$. *Hence, defining* $\mathcal{R} := \frac{\Delta\mathcal{L}^\star_{\mathcal{S}_2}}{\Delta\mathcal{L}^\star_{\mathcal{S}_\infty}} = \frac{\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\|_F^4}{\|\boldsymbol{G}\boldsymbol{K}^{-1}\mathcal{A}\|_F^2}\frac{\|\boldsymbol{B}\mathcal{A}\|_F^2}{\langle \boldsymbol{G},\boldsymbol{B}\rangle^2}$, *the $\mathcal{S}_2$ constraint achieves a larger optimal decrease if $\mathcal{R} \geq 1$, while the $\mathcal{S}_\infty$ constraint achieves a larger optimal decrease if $\mathcal{R} < 1$.*

By Theorem 2, the value of $\mathcal{R}$ determines which norm constraint should be used: when $\mathcal{R} \geq 1$, the $\mathcal{S}_2$ norm constraint is preferred, whereas when $\mathcal{R} < 1$, the $\mathcal{S}_\infty$ norm constraint is more appropriate. We put the proof of Theorem 2 in Appendix A.2. This framework can be naturally extended to a broader range of choices for p, and may further lead to algorithms with layer-wise and time-dependent p (Wu et al., 2026; Shumaylov et al., 2026; Massena et al., 2026). We leave these extensions to future work.

Here, we consider an adaptive variant of SAFESCHATTEN on the Qwen-7B-Instruct model on AG News dataset based on Theorem 2. Specifically, we use $\mathcal{R}$ to determine whether to apply SAFESCHATTEN−2 or SAFESCHATTEN−∞ at the current step. For computational efficiency, we update the choice of p every five steps. For all three variants SAFESCHATTEN−2, SAFESCHATTEN−∞, and the adaptive method, we fine-tune only the Q and K projection layers.

As shown in Table 6, the SAFESCHATTEN-Adaptive achieves higher downstream-task accuracy than both SAFESCHATTEN−2 and SAFESCHATTEN−∞, while maintaining competitive harmfulness scores. These empirical results are consistent with the theoretical prediction of Theorem 2. In Figure 4, we visualize the fraction of layers that select SAFESCHATTEN−2 or SAFESCHATTEN−∞ at each training step. We observe a clear trend: SAFESCHATTEN−2 is selected more frequently during the early stages of training, whereas SAFESCHATTEN−∞ becomes increasingly preferred at later stages. Interestingly, this trend is also consistent with observations from prior work (Davis & Drusvyatskiy, 2025; Massena et al., 2026; Wu et al., 2026) studying when Muon outperforms gradient descent.

4 CONCLUSION

This work introduces SAFESCHATTEN, which approaches safety-preserving fine-tuning by asking how much a task update may change protected activations. Schatten-norm budgets induce distinct update directions across singular modes. Both Frobenius and spectral constraints reduce harmfulness while preserving downstream adaptation, supporting activation-constrained optimization as a practical defense and motivate training-wide budget policies and broader behavioral evaluation.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing, as well as for assisting with the development of code for result visualization. We did not use generative AI tools to generate research ideas or to write proofs of lemmas or theorems, and the remaining required disclosure categories are not applicable to this work. We have reviewed all AI-assisted work. In particular, we carefully checked AI-polished text to ensure that it accurately preserved our intended meaning, and LLM-generated code was verified and tested for correctness by two authors. We take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to reduce the loss of safety alignment during language-model fine-tuning. Our experiments use existing benchmark datasets and pretrained models, including harmful examples from BeaverTails to evaluate safety degradation and defenses. Studying how fine-tuning weakens safeguards carries a dual-use risk: the resulting insights could also inform attempts to bypass alignment. We use these experiments to develop and assess protective optimization methods. Our safety evaluation relies on a finite set of harmful prompts and an automated moderation model, whose judgments may contain errors and biases and do not capture all dimensions of safety, fairness, or privacy. Consequently, lower measured harmfulness and bounded activation drift do not establish a general guarantee of safe behavior, particularly under unseen attacks or distribution shifts. We view SAFESCHATTEN as one component of a broader safety process that includes application-specific evaluation and safeguards.

REPRODUCIBILITY STATEMENT

Our code is available at https://anonymous.4open.science/r/safeschatten-9527/. Section 2 describes the activation constraints, theoretical assumptions, and update rules of SAFESCHATTEN, with pseudocode in Algorithm 1. Section 3.4.3 presents the criterion for adaptive norm selection, and Appendix A.2 provides the proofs. Section 3.1 identifies the datasets, models, baselines, harmful-data mixing procedure, and evaluation metrics. Appendix A.3 documents the training hyperparameters, random seed, numerical precision, safety-activation cache construction, and normalization and values of the update budgets for the reported task configurations. The matrix-sign, momentum, and cache-refresh ablations are described in Sections 3.4.1 and 3.4.2; Appendix A.4 provides the data-size and harmful-fraction sweeps. As noted in Appendix A.3, the reported results are individual-run measurements, and small differences should not be interpreted as statistically established improvements.

REFERENCES

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, 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. URL https://openreview.net/forum?id=gT5hALch9z.

Pin-Yu Chen, Han Shen, Payel Das, and Tianyi Chen. Fundamental safety-capability trade-offs in fine-tuning large language models. *PNAS Nexus*, 5(4):pgag097, 04 2026.

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.

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M. Gower. An exploration of non-euclidean gradient descent: Muon and its many variants, 2025. URL https://arxiv.org/abs/2510.09827.

Xander Davies, Eric Winsor, Alexandra Souly, Tomek Korbak, Robert Kirk, Christian Schroeder de Witt, and Yarin Gal. Fundamental limitations in pointwise defences of llm finetuning apis. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume

38, Main Conference, pp. 75931–75987. Curran Associates, Inc., 2025. doi: 10.52202/085713-2551. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6ded182245f8e4db3963d2027174133d-Paper-Conference.pdf.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. Rmnp: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

GemmaTeam, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. Memory-efficient llm pretraining via minimalist optimizer design. *arXiv preprint arXiv:2506.16659*, 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.

Yufei Gu and Zeke Xie. Mano: Restriking manifold optimization for llm training. *arXiv preprint arXiv:2601.23000*, 2026.

Luxi He, Mengzhou Xia, and Peter Henderson. What is in Your Safe Data? Identifying Benign Data that Breaks Safety. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Hi8jKh4HE9.

Lei Hsiung, Tianyu Pang, Yung-Chen Tang, Linyue Song, Tsung-Yi Ho, Pin-Yu Chen, and Yaoqing Yang. Why LLM safety guardrails collapse after fine-tuning: A similarity analysis between

alignment and fine-tuning datasets. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16601–16619, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.756. URL https://aclanthology.org/2026.acl-long.756/.

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, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack. *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RPChapuXlC.

Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning, 2025a. URL https://arxiv.org/abs/2408.09600.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating Harmful Perturbation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=tTPHgb0EtV.

Jiaming Ji, Mickel Liu, Juntao 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Kaizhao Liang, Lizhang Chen, Bo Liu, and Qiang Liu. Cautious optimizers: Improving training with one line of code. In *International Conference on Learning Representations*, volume 2026, pp. 106538–106563, 2026.

Guozhi Liu, Weiwei Lin, Tiansheng Huang, Ruichao Mo, Qi Mu, Xiumin Wang, and Li Shen. Surgery: Mitigating harmful fine-tuning for large language models via attention sink. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=6ojsIliNF0.

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. URL https://arxiv.org/abs/2505.12038.

Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. Swan: Sgd with normalization and whitening enables stateless llm training. *arXiv preprint arXiv:2412.13148*, 2024.

Thomas Massena, Corentin Friedrich, and Mathieu Serrurier. From sgd to muon: Adaptive optimization via schatten-p norms. *arXiv preprint arXiv:2605.19781*, 2026.

Quoc Nguyen, Trung Le, Jing Wu, Anh Bui, and Mehrtash Harandi. Antibody: Strengthening defense against harmful fine-tuning for large language models via attenuating harmful gradient influence. In *International Conference on Learning Representations*, volume 2026, pp. 121713–121741, 2026.

Tianyu Pang, Yujie Fang, Zihang Liu, Shenyang Deng, Lei Hsiung, Shuhua Yu, and Yaoqing Yang. Htmuon: Improving muon via heavy-tailed spectral correction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36504–36535, 2026.

ShengYun Peng, Pin-Yu Chen, Jianfeng Chi, Seongmin Lee, and Duen Horng Chau. Shape it up! restoring LLM safety during finetuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PAIVwOaAnq.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos, 2025. URL https://arxiv.org/abs/2502.07529.

Kaustubh Ponkshe, Shaan Shah, Raghav Singhal, and Praneeth Vepakomma. Safety subspaces are not linearly distinct: A fine-tuning case study. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Fj6LakRHcT.

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 *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making muon & scion great again! (bridging theory and practice of lmo-based optimizers for llms), 2025. URL https://arxiv.org/abs/2505.13416.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight llms. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 101802–101829, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fc49a629d33bc2461ed7a715ce44da68-Paper-Conference.pdf.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why muon outperforms adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026.

Ziqing Wen, Yanqi Shi, Jiahuan Wang, Ping Luo, Linbo Qiao, Dongsheng Li, and Tao Sun. Sron: State-free llm training via row-wise gradient normalization. 2025.

Fangzhou Wu, Rikhav Shah, Sandeep Silwal, and Qiuyi Zhang. Dynmuon: A dynamic spectral shaping view of muon. *arXiv preprint arXiv:2605.17109*, 2026.

Zhixin Xie, Xurui Song, and Jun Luo. Attack via overfitting: 10-shot benign fine-tuning to jailbreak llms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 31136–31170. Curran Associates, Inc., 2025. doi: 10.52202/085713-1046. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2cb880950081ceb85100951b0ad0d542-Paper-Conference.pdf.

Ruihan Xu, Jiajin Li, and Yiping Lu. On the width scaling of neural optimizers under matrix operator norms i: Row/column normalization and hyperparameter transfer. *arXiv preprint arXiv:2603.09952*, 2026.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Shuo Yang, Qihui Zhang, Yuyang Liu, Yue Huang, Xiaojun Jia, Kun-Peng Ning, Jia-Yu Yao, Jigang Wang, Dai Hailiang, Yibing Song, and Li Yuan. Asft: Anchoring safety during llm fine-tuning within narrow safety basin. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34322–34330, 2026.

Bingjie Zhang, Yibo Yang, Ren Zhe, Dandan Guo, Jindong Gu, Philip Torr, and Bernard Ghanem. A guardrail for safety preservation: When safety-sensitive subspace meets harmful-resistant null-space. In *International Conference on Learning Representations*, volume 2026, pp. 134322–134347, 2026.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 649–657, Cambridge, MA, USA, 2015. MIT Press.

A Appendix

A.1 Related Work

Norm-Constrained Steepest Descent. Recent work (Pethick et al., 2025) has shown that a broad class of first-order optimizers can be interpreted as steepest-descent methods through the lens of linear minimization oracles (LMOs). This perspective provides a unified framework for designing optimization updates under a variety of non-Euclidean geometries. For example, directly constraining the update matrix by a Schatten-∞ norm recovers the matrix-sign update underlying Muon (Jordan et al., 2024), which has recently gained widespread adoption in large-scale LLM pretraining. More generally, imposing Schatten-p constraints gives rise to Muon-style variants such as HTMuon (Pang et al., 2026), while max row- or column-wise norm constraints lead to a family of recent row/column-normalized optimizers, including RMNP (Deng et al., 2026), SRON (Wen et al., 2025), SCAL (Glentis et al., 2025), SWAN (Ma et al., 2024), MANO (Gu & Xie, 2026), and MOGA (Xu et al., 2026). Likewise, an element-wise ℓ_∞ constraint yields the SignGD (Liang et al., 2026) update. The LMO perspective can be further extended from individual matrices to layer-wise optimization geometries, as explored in methods such as Gluon (Riabinin et al., 2025) and Momo (Crawshaw et al., 2025).

In our work, we impose a Schatten-p norm constraint on safety-relevant activation changes and derive the corresponding steepest-descent update. Our key premise is that not only the overall magnitude of activation drift, but also how this drift is distributed across singular directions, is important for preserving safety during fine-tuning. By explicitly accounting for the spectral geometry of activation drift, this steepest-descent update provides a principled way to limit harmful changes along safety-relevant directions while still allowing effective adaptation to the downstream task, thereby better preserving the model's safety capabilities during fine-tuning.

Safety preservation during fine-tuning. Safety degradation can arise under both adversarial and benign customization (Qi et al., 2024; He et al., 2024; Xie et al., 2025; Hsiung et al., 2026). Existing defenses intervene at several stages: strengthening a model before adaptation (Huang et al., 2025b; Tamirisa et al., 2025), incorporating safety supervision during training (Bianchi et al., 2024; Huang et al., 2024), modifying gradients or update subspaces (Nguyen et al., 2026; Hsu et al., 2024; Zhang et al., 2026), or correcting parameter changes afterward (Lu et al., 2025; Huang et al., 2025a). Activation-based defenses are particularly relevant because they connect a parameter update to its effect on safety-triggering inputs. In particular, Zhang et al. (2026) protects safety-sensitive weight components and restricts adapter updates using a null-space projector. In contrast, SafeSchatten permits a finite amount of activation drift while selecting an update direction that optimizes the downstream task objective subject to a chosen Schatten-norm budget. The budget controls the overall amount of permitted activation change, while the choice of norm controls how that change is distributed across singular directions.

A.2 Theoretical Analysis of SafeSchatten

Lemma 2 (Von Neumann trace inequality)**.** *For matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, let $\sigma_i(\boldsymbol{A})$ denote the i-th largest singular value of $\boldsymbol{A}$, and let $r = \min(m, n)$. Then for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, we have*

$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{Tr}\left(\boldsymbol{A}^\top \boldsymbol{B}\right) \le \sum_{i=1}^{r} \sigma_i(\boldsymbol{A})\, \sigma_i(\boldsymbol{B}).$$

Lemma 3. *Let $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the SVD of the negative gradient $\boldsymbol{G} = -\nabla f(\boldsymbol{W})$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Let p be the conjugate exponent of q, satisfying $1/p + 1/q = 1$. The explicit solution $\Delta\boldsymbol{W}^*$ to the optimization problem 6 is given by:*

$$\Delta\boldsymbol{W}^* = -\eta\varepsilon \frac{1}{\|\boldsymbol{G}\|_{\mathcal{S}_q}^{q/p}} \boldsymbol{U}\boldsymbol{\Sigma}^{q-1}\boldsymbol{V}^\top, \tag{14}$$

where $\boldsymbol{\Sigma}^{q-1} = \mathrm{diag}(\sigma_1^{q-1}, \ldots, \sigma_r^{q-1})$.

Proof. First, by invoking Von Neumann's trace inequality in Lemma 2, we have the upper bound

$$\mathrm{Tr}(\boldsymbol{G}^\top \Delta\boldsymbol{W}) \le \sum_i \sigma_i(\boldsymbol{G})\sigma_i(\Delta\boldsymbol{W})$$

The equality holds if and only if $\Delta \boldsymbol{W}$ shares the same left and right singular vectors as $\boldsymbol{G}$. Thus, the optimal $\Delta \boldsymbol{W}$ must take the form:

$$\Delta \boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top, \quad \text{where } \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_r). \tag{15}$$

Here, $\lambda_i \geq 0$ represents the singular values of $\Delta \boldsymbol{W}$. The problem reduces to maximizing $\sum_i \sigma_i \lambda_i$ subject to $\|\lambda\|_p \leq \delta$. By Hölder's inequality, we have the upper bound:

$$\sum_i \sigma_i \lambda_i \leq \|\sigma\|_q \|\lambda\|_p \leq \|\sigma\|_q \eta\varepsilon. \tag{16}$$

Equality is achieved if and only if $\lambda_i^p \propto \sigma_i^q$, which implies:

$$\lambda_i = c\sigma_i^{q/p}, \tag{17}$$

for some constant $c \geq 0$. Using the conjugate property $1/q + 1/p = 1$, the exponent simplifies to:

$$\frac{q}{p} = q\left(1 - \frac{1}{q}\right) = q - 1. \tag{18}$$

Substituting this back yields the optimal singular values:

$$\lambda_i = c \cdot \sigma_i^{q-1}, \tag{19}$$

where c is a normalization constant. To determine c, we substitute these values into the active constraint $\|\Delta \boldsymbol{W}\|_p = \eta\varepsilon$:

$$\left(\sum_i (c\sigma_i^{q-1})^p\right)^{1/p} = c\left(\sum_i \sigma_i^{(q-1)p}\right)^{1/p} = \eta\varepsilon. \tag{20}$$

Using the algebraic relation $(q-1)p = q$, the term involving singular values simplifies to the Schatten$-q$ norm of $\boldsymbol{G}$. Solving for c, we obtain:

$$c\|\boldsymbol{G}\|_{\mathcal{S}_q}^{q/p} = \delta \implies c = \frac{\eta\varepsilon}{\|\boldsymbol{G}\|_{\mathcal{S}_q}^{q/p}}. \tag{21}$$

Substituting c back into the expression $\Delta \boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ yields the stated closed-form solution. □

Theorem 3. *Let $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top$ defined as the safety-activation covariance. Assume that $\boldsymbol{K} \succ \boldsymbol{0}$. Then, for $p \in \{2, \infty\}$, under Assumption 1, the layer-wise steepest descent directions for optimization problem (10) are given by*

$$\boldsymbol{\Phi}^\star = \begin{cases} \varepsilon \dfrac{\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}}{\left\|\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}\right\|_F}\boldsymbol{K}^{-1/2}, & p = 2, \\ \varepsilon \operatorname{msign}\left(\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}\right)\boldsymbol{K}^{-1/2}, & p = \infty. \end{cases} \tag{22}$$

for an SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, we define $\operatorname{msign}(\boldsymbol{A}) := \boldsymbol{U}\boldsymbol{V}^\top$.

Proof. We consider first-order approximation for $\mathcal{A}(\boldsymbol{X}, \boldsymbol{W} + \Delta \boldsymbol{W}) - \mathcal{A}(\boldsymbol{X}, \boldsymbol{W})$, we have

$$\mathcal{A}(\boldsymbol{X}, \boldsymbol{W} + \Delta \boldsymbol{W}) - \mathcal{A}(\boldsymbol{X}, \boldsymbol{W}) \approx -\eta\sigma'(\boldsymbol{W}\boldsymbol{X}) \odot (\boldsymbol{\Phi}\boldsymbol{X}),$$

by Assumption 1, we can translate the problem in (10) to

$$\min_{\boldsymbol{\Phi}} \; Tr(\boldsymbol{G}^\top \boldsymbol{\Phi}) \quad \text{s.t.} \quad \|\boldsymbol{\Phi}\boldsymbol{X}\|_{\mathcal{S}_p} \leq \varepsilon \tag{23}$$

Assume $\widetilde{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\boldsymbol{X}$, and set $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top$, we have

$$\min_{\widetilde{\boldsymbol{\Phi}}} \; Tr((\boldsymbol{G}\boldsymbol{K}^{-\frac{1}{2}})^\top \widetilde{\boldsymbol{\Phi}}) \quad \text{s.t.} \quad \|\widetilde{\boldsymbol{\Phi}}\|_{\mathcal{S}_p} \leq \varepsilon. \tag{24}$$

By Lemma 1, the solution will be

$$\boldsymbol{\Phi}^\star = \begin{cases} \varepsilon \dfrac{\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}}{\left\|\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}\right\|_F}\boldsymbol{K}^{-1/2}, & p = 2, \\ \varepsilon \operatorname{msign}\left(\boldsymbol{G}_{\text{task}}\boldsymbol{K}^{-1/2}\right)\boldsymbol{K}^{-1/2}, & p = \infty. \end{cases} \tag{25}$$

□

Theorem 4 (Choosing between $\mathcal{S}_2$ and $\mathcal{S}_\infty$ constraints)**.** *Suppose we consider a simplified surrogate loss in (13) Let* $\boldsymbol{B} := \operatorname{msign}\big(\boldsymbol{G}\boldsymbol{K}^{-1/2}\big)\,\boldsymbol{K}^{-1/2}$. *For the optimal* $\mathcal{S}_2$ *and* $\mathcal{S}_\infty$ *update directions, optimizing over the step size* η *gives*

$$\Delta\mathcal{L}^\star_{\mathcal{S}_2} = \frac{N}{2}\frac{\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\|_F^4}{\|\boldsymbol{G}\boldsymbol{K}^{-1}\mathcal{A}\|_F^2}, \qquad \Delta\mathcal{L}^\star_{\mathcal{S}_\infty} = \frac{N}{2}\frac{\langle \boldsymbol{G}, \boldsymbol{B}\rangle^2}{\|\boldsymbol{B}\mathcal{A}\|_F^2}.$$

Hence, defining

$$\mathcal{R} := \frac{\Delta\mathcal{L}^\star_{\mathcal{S}_2}}{\Delta\mathcal{L}^\star_{\mathcal{S}_\infty}} = \frac{\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\|_F^4}{\|\boldsymbol{G}\boldsymbol{K}^{-1}\mathcal{A}\|_F^2}\frac{\|\boldsymbol{B}\mathcal{A}\|_F^2}{\langle \boldsymbol{G}, \boldsymbol{B}\rangle^2},$$

the $\mathcal{S}_2$ *constraint achieves a larger optimal decrease if* $\mathcal{R} \geq 1$*, while the* $\mathcal{S}_\infty$ *constraint achieves a larger optimal decrease if* $\mathcal{R} < 1$.

Proof.

Case 1: $p = 2$**.** By Theorem 1, we have $\Delta\boldsymbol{W} = -\eta\boldsymbol{\Phi} = -\eta\varepsilon\frac{\boldsymbol{G}\boldsymbol{K}^{-1}}{\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\|_F}$. Plugging this into $\mathcal{L}(\boldsymbol{W} + \Delta\boldsymbol{W}) = \mathcal{L}(\boldsymbol{W}) + \langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle + \frac{1}{2N}\|\Delta\boldsymbol{W}\mathcal{A}\|_F^2$, we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{W} - \eta\boldsymbol{\Phi}) &= \mathcal{L}(\boldsymbol{W}) - \eta\langle \boldsymbol{G}, \boldsymbol{\Phi}\rangle + \frac{\eta^2}{2N}\|\boldsymbol{\Phi}\mathcal{A}\|_F^2 \\ &= \mathcal{L}(\boldsymbol{W}) - \eta\varepsilon\left\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\right\|_F + \frac{\eta^2}{2N}\|\boldsymbol{\Phi}\mathcal{A}\|_F^2.\end{aligned}$$

So the $\Delta\mathcal{L}^\star_{\mathcal{S}_2} = |\mathcal{L}(\boldsymbol{W} - \eta\boldsymbol{\Phi}) - \mathcal{L}(\boldsymbol{W})|$ is the quadratic function of η. To maximize the guaranteed decrease, we choose $\eta \sim \frac{N\varepsilon\left\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\right\|_F}{\|\boldsymbol{\Phi}\mathcal{A}\|_F^2}$, so for $p = 2$, the guaranteed decrease is approximately

$$\Delta\mathcal{L}^\star_{\mathcal{S}_2} \sim \frac{1}{2}N\varepsilon^2\frac{\left\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\right\|_F^2}{\|\boldsymbol{\Phi}\mathcal{A}\|_F^2} = \frac{1}{2}N\frac{\left\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\right\|_F^4}{\|\boldsymbol{G}\boldsymbol{K}^{-1}\mathcal{A}\|_F^2}.$$

Case 2: $p = \infty$**.** By Theorem 1, we have $\Delta\boldsymbol{W} = -\eta\boldsymbol{\Phi} = -\eta\varepsilon\operatorname{msign}\big(\boldsymbol{G}\boldsymbol{K}^{-1/2}\big)\,\boldsymbol{K}^{-1/2}$. Plugging this into $\mathcal{L}(\boldsymbol{W} + \Delta\boldsymbol{W}) = \mathcal{L}(\boldsymbol{W}) + \langle \boldsymbol{G}, \Delta\boldsymbol{W}\rangle + \frac{1}{2N}\|\Delta\boldsymbol{W}\mathcal{A}\|_F^2$, we have

$$\mathcal{L}(\boldsymbol{W} - \eta\boldsymbol{\Phi}) = \mathcal{L}(\boldsymbol{W}) - \eta\langle \boldsymbol{G}, \boldsymbol{\Phi}\rangle + \frac{\eta^2}{2N}\|\boldsymbol{\Phi}\mathcal{A}\|_F^2$$

So the $\Delta\mathcal{L}^\star_{\mathcal{S}_\infty} = |\mathcal{L}(\boldsymbol{W} - \eta\boldsymbol{\Phi}) - \mathcal{L}(\boldsymbol{W})|$ is the quadratic function of η. To maximize the guaranteed decrease, we choose $\eta \sim \frac{N\langle \boldsymbol{G}, \boldsymbol{\Phi}\rangle}{\|\boldsymbol{\Phi}\mathcal{A}\|_F^2}$, so for $p = \infty$, the guaranteed decrease is approximately

$$\Delta\mathcal{L}^\star_{\mathcal{S}_\infty} \sim \frac{1}{2}N\frac{\langle \boldsymbol{G}, \boldsymbol{\Phi}\rangle^2}{\|\boldsymbol{\Phi}\mathcal{A}\|_F^2} = \frac{1}{2}N\frac{\langle \boldsymbol{G}, \operatorname{msign}\big(\boldsymbol{G}\boldsymbol{K}^{-1/2}\big)\,\boldsymbol{K}^{-1/2}\rangle^2}{\|\operatorname{msign}\big(\boldsymbol{G}\boldsymbol{K}^{-1/2}\big)\,\boldsymbol{K}^{-1/2}\mathcal{A}\|_F^2}.$$

Since we assume $\boldsymbol{B} := \operatorname{msign}\big(\boldsymbol{G}\boldsymbol{K}^{-1/2}\big)\,\boldsymbol{K}^{-1/2}$., so the value of $\frac{\Delta\mathcal{L}^\star_{\mathcal{S}_2}}{\Delta\mathcal{L}^\star_{\mathcal{S}_\infty}} = \frac{\|\boldsymbol{G}\boldsymbol{K}^{-1/2}\|_F^4}{\|\boldsymbol{G}\boldsymbol{K}^{-1}\mathcal{A}\|_F^2}\frac{\|\boldsymbol{B}\mathcal{A}\|_F^2}{\langle \boldsymbol{G}, \boldsymbol{B}\rangle^2} = \mathcal{R}$ determines which norm constraint should be used: when $\frac{\Delta\mathcal{L}^\star_{\mathcal{S}_2}}{\Delta\mathcal{L}^\star_{\mathcal{S}_\infty}} \geq 1$, the $\mathcal{S}_2$ norm constraint is preferred, whereas when $\frac{\Delta\mathcal{L}^\star_{\mathcal{S}_2}}{\Delta\mathcal{L}^\star_{\mathcal{S}_\infty}} < 1$, the $\mathcal{S}_\infty$ norm constraint is more appropriate. □

A.3 Training details.

The reported SafeSchatten runs use rank-8 LoRA adapters on the query and value projections of all 28 decoder blocks, with zero adapter dropout. We train for 10 epochs with batch size 8, sequence length 1024, seed 42, and no weight decay. The learning-rate schedule uses 10% linear warmup followed by linear decay, with peak rates 10^{-5}, 2×10^{-5}, and 5×10^{-5} for SST2, AG News, and GSM8K, respectively. Model computation uses bfloat16 and the safety solves use float32. We cache

Table 7: Safety and task performance after fine-tuning Llama-3.2-3B-Instruct on AG News with different number of training data n

	Harm (↓)					Acc. (↑)					$\|\text{SAD}\|_F$					$\|\text{SAD}\|$				
num of training data (n)	0.5K	1.0K	1.5K	2.0K	2.5K	0.5K	1.0K	1.5K	2.0K	2.5K	0.5K	1.0K	1.5K	2.0K	2.5K	0.5K	1.0K	1.5K	2.0K	2.5K
Base Model			19.80					53.37					-					-		
LoRA	47.80	51.90	54.60	54.90	54.60	60.83	87.08	86.99	89.76	87.89	0.3894	0.4278	0.4461	0.4549	0.4638	0.2509	0.2917	0.3159	0.3284	0.3315
SafeDelta	26.40	35.00	37.70	35.20	32.40	61.72	88.38	90.11	90.13	89.34	0.4418	0.4418	0.4307	0.4205	0.4120	0.3532	0.3140	0.3003	0.2736	0.2692
DSS	12.40	25.40	14.90	24.50	35.20	57.22	87.68	88.17	90.29	89.99	1.8407	1.6528	1.5902	1.5984	1.3850	2.3463	2.0783	1.9746	2.0431	1.7214
Surgery	11.80	36.70	21.80	12.50	17.50	24.20	86.28	53.47	76.24	90.87	0.7094	0.7480	0.7639	0.7493	0.7766	0.4281	0.4737	0.4353	0.4559	0.4595
Safe LoRA	25.40	25.20	22.40	23.00	22.70	73.86	78.99	79.18	79.47	79.09	0.1167	0.1320	0.1505	0.1594	0.1541	0.0792	0.0791	0.0852	0.0870	0.0852
AsFT	22.50	21.30	21.70	22.10	23.00	73.90	76.50	77.10	75.10	74.80	0.1037	0.1125	0.1339	0.1251	0.1394	0.0579	0.0552	0.0629	0.0586	0.0669
GuardSpace	11.00	20.10	24.30	24.40	27.20	56.90	88.90	88.30	87.80	87.50	0.4061	0.4421	0.3969	0.4484	0.4354	0.2608	0.2948	0.2478	0.3251	0.3195
SAFESCHATTEN−2	11.80	15.70	20.00	25.90	23.90	56.20	84.30	84.97	86.59	86.00	0.1603	0.2811	0.3135	0.3395	0.3333	0.1088	0.2024	0.2339	0.2568	0.2525
SAFESCHATTEN−∞	13.00	16.20	24.40	33.50	28.90	55.62	85.93	86.97	87.50	87.64	0.1081	0.2704	0.3255	0.3660	0.3587	0.0680	0.1982	0.2499	0.2831	0.2807

Table 8: Safety and task performance after fine-tuning Llama-3.2-3B-Instruct on AG News with different harmful fraction ρ. Clean denotes $\rho = 0$

	Harm (↓)					Acc. (↑)					$\|\text{SAD}\|_F$					$\|\text{SAD}\|$				
harmful fraction (ρ)	clean	0.05	0.10	0.15	0.20	clean	0.05	0.10	0.15	0.20	clean	0.05	0.10	0.15	0.20	clean	0.05	0.10	0.15	0.20
Base Model			19.80					53.37					-					-		
LoRA	19.70	47.90	51.90	53.80	55.70	87.14	86.47	87.08	86.67	86.87	0.1066	0.3858	0.4278	0.4319	0.4628	0.0507	0.2520	0.2917	0.2985	0.3247
SafeDelta	17.40	26.30	35.00	41.90	37.60	88.74	87.93	88.38	89.07	88.53	0.5257	0.4856	0.4418	0.4001	0.4747	0.4983	0.4110	0.3140	0.2556	0.3677
DSS	9.20	17.70	25.40	23.30	20.30	88.03	88.30	87.68	87.80	87.54	1.5861	1.6194	1.6528	1.5457	1.4386	1.9800	2.0292	2.0783	1.9424	1.8064
Surgery	3.50	27.20	36.70	15.30	27.00	81.97	85.25	86.28	64.45	56.13	1.0477	0.7586	0.7480	0.7479	0.7486	1.0997	0.4800	0.4737	0.4571	0.4451
Safe LoRA	18.70	21.60	25.20	25.20	25.30	76.39	77.54	78.99	79.71	80.34	0.0837	0.1170	0.1320	0.1443	0.1615	0.0398	0.0691	0.0791	0.0858	0.0937
AsFT	19.30	22.10	21.30	23.20	21.90	75.30	78.60	76.50	76.60	72.90	0.0676	0.1018	0.1125	0.1241	0.1409	0.0286	0.0510	0.0552	0.0618	0.0659
GuardSpace	9.20	18.80	20.10	23.40	24.00	88.60	87.80	88.90	88.10	87.50	0.1753	0.3998	0.4421	0.4292	0.4542	0.0939	0.2655	0.2948	0.2928	0.3260
SAFESCHATTEN−2	16.30	15.20	15.70	18.70	24.70	85.13	83.92	84.30	84.03	83.42	0.0532	0.1814	0.2811	0.3160	0.3359	0.0328	0.1199	0.2024	0.2330	0.2514
SAFESCHATTEN−∞	15.20	12.10	16.20	22.60	28.80	86.24	85.79	85.93	83.86	83.86	0.0328	0.1284	0.2704	0.3174	0.3556	0.0157	0.0838	0.1982	0.2417	0.2729

activations from 100 harmful prompts and the aligned model's continuations, capped at 64 generated tokens per prompt and 100,000 cached tokens per target matrix.

The implementation rescales each candidate direction to a fixed measured update budget. Frobenius drift is normalized by the square root of the number of drift-matrix elements; spectral drift is normalized by the square root of the number of cached tokens. In SST2/AG News/GSM8K order, the fixed step budgets are $(2 \times 10^{-6}, 6 \times 10^{-6}, 10^{-7})$ for SAFESCHATTEN-2 and $(4 \times 10^{-5}, 6 \times 10^{-5}, 10^{-7})$ for SAFESCHATTEN-∞. These are budgets on the realized step, corresponding to $\eta\varepsilon$ in Section 2, and have different normalization conventions across the two geometries. The AG News sweeps retain the corresponding main-run budget. Results are individual-run measurements; small differences should not be interpreted as statistically established improvements.

A.4 Sensitivity to Data Size and Poisoning

We test whether the safety gains persist as the scale and composition of fine-tuning data change. On AG News with Llama-3.2-3B-Instruct, we vary the dataset size from 0.5K to 2.5K at a fixed harmful fraction $\rho = 0.10$, and separately vary ρ from 0 to 0.20 at a fixed size of 1,000 examples. SAFESCHATTEN−2 and SAFESCHATTEN−∞ retains its main-experiment budget throughout both sweeps. Training lasts ten epochs, so the size sweep increases both the number of optimizer updates and exposure to harmful examples. Tables 7 and 8 provide the complete measurements, including the relative activation-drift diagnostics defined in Eq. 1.

Safety gains persist as the training set grows. SAFESCHATTEN−2 and SAFESCHATTEN−∞ remain less harmful than LoRA at every tested dataset size while benefiting from additional task data. At 2.5K examples, SAFESCHATTEN-2 achieves 86.00% accuracy with 23.90% harmfulness, compared with 87.89% and 54.60% for LoRA: a 30.70-percentage-point reduction in harmfulness for a 1.89-point accuracy cost. SAFESCHATTEN−∞ offers another favorable operating point, retaining 87.64% accuracy with 28.90% harmfulness. These results show that the safety advantage extends to larger training sets while retaining most of the task performance of standard fine-tuning.

Activation constraints mitigate harmful-data contamination. SAFESCHATTEN−2 and SAFESCHATTEN−∞ also remain less harmful than LoRA across all tested harmful fractions. At $\rho = 0.20$, SAFESCHATTEN-2 reduces harmfulness from LoRA's 55.70% to 24.70%, while retaining 83.42% accuracy versus 86.87%. This yields a 31.00-point harmfulness reduction for a 3.45-point accuracy cost. Its accuracy remains between 83.42% and 85.13% throughout the sweep, demonstrating useful task adaptation even as harmful examples replace a larger share of the training data.

SAFESCHATTEN$-\infty$ exhibits the same safety advantage over LoRA, reaching 83.86% accuracy with 28.80% harmfulness at the highest harmful fraction.

Budget selection should account for training exposure. The fixed budgets preserve a substantial safety advantage over LoRA in both sweeps. At the same time, increasing the dataset size from 0.5K to 2.5K raises SAFESCHATTEN-2's relative Frobenius drift from 0.1603 to 0.3333 and harmfulness from 11.80% to 23.90% (Table 7). This links the observed tradeoff to cumulative training exposure: a per-step constraint limits each update, while a longer sequence of updates can accumulate representational change. The results motivate choosing the activation budget together with the training duration and harmful-data exposure to maintain the desired safety-utility balance.