Curvature-Aware Safety Restoration In LLMs Fine-Tuning

Anonymous authors
Paper under double-blind review

Abstract

Fine-tuning Large Language Models (LLMs) for downstream tasks often compromises safety alignment, even when using parameter-efficient methods like LoRA. In this work, we uncover a notable property: fine-tuned models preserve the geometric structure of their loss landscapes concerning harmful content, regardless of the fine-tuning method employed. This suggests that safety behaviors are not erased but shifted to less influential regions of the parameter space. Building on this insight, we propose a curvature-aware alignment restoration method that leverages influence functions and second-order optimization to selectively increase loss on harmful inputs while preserving task performance. By navigating the shared geometry between base and fine-tuned models, our method discourages unsafe outputs while preserving task-relevant performance, avoiding full reversion and enabling precise, low-impact updates. Extensive evaluations across multiple model families and adversarial settings show that our approach efficiently reduces harmful responses while maintaining or even improving utility and few-shot learning performance.

1 Introduction

Large Language Models (LLMs) encode safety-aligned behaviors during pretraining, but these safeguards deteriorate during task-specific fine-tuning, a phenomenon we identify as *safety alignment drift*. Studies demonstrate that even minimal fine-tuning can compromise safety mechanisms, with models like GPT-3.5 Turbo becoming consistently unsafe after adaptation on just 10 adversarial examples Qi et al. (2023b). Attempts to address this issue by modifying model behavior generally fall into two main categories, both of which suffer from inherent limitations. **Behavioral unlearning** methods attempt to remove undesirable knowledge or responses (Cao & Yang, 2015; Bourtoule et al., 2021a), but often require costly retraining or risk catastrophic forgetting. **Model editing** approaches aim to update factual associations or local behaviors through direct parameter intervention (Meng et al., 2022; Mitchell et al., 2022), yet struggle to generalize beyond narrow scopes or isolated prompts. To solve these issues, we propose a new direction that treats safety behavior as an intrinsic property of the model's geometry and seeks to restore alignment through curvature-aware navigation of the loss landscape.

Our key insight, supported by extensive empirical analysis (Section 2), is that models preserve notable structural properties in their loss landscapes with respect to harmful content after fine-tuning. Specifically, we observe high correlations in models' responses to harmful inputs before and

Table 1: Pearson correlation coefficients between base and fine-tuned models' responses on harmful content (HEx-PHI), task-specific data (Dolly), and general data (Alpaca). High correlations (>0.77) for harmful content across all models indicate preserved safety structure, while low correlations on task/general data show significant behavioral changes during fine-tuning. This asymmetric preservation validates our hypothesis that safety mechanisms remain structurally intact in the loss landscape.

Models	**Harmful**	**Dolly**	**Alpaca**
LLama-2 7B	**0.992**	0.056	-0.055
LLama-2 13B	**0.994**	0.084	0.085
LLama 3 8B	**0.995**	0.550	0.510
Mistra v3 7B	**0.771**	0.167	0.087
Gemma 2	**0.799**	0.291	0.199
Qwen 2.5 7B	**0.994**	0.014	0.067

after fine-tuning, despite substantial divergence in other functional behaviors. This suggests that safety mechanisms remain largely preserved in the parameter space, merely shifted to less dominant regions during task-specific optimization.

This observation motivates our novel approach: *curvature-aware alignment restoration*. We leverage the preserved geometry of the loss landscape to restore safety boundaries. By employing influence functions and second-order optimization techniques, our method navigates the parameter space to increase loss on harmful inputs while minimizing impact on task performance. Our contributions include:

- We identify and empirically validate a key insight: Fine-tuning preserves the geometric structure of the loss landscape for harmful content across diverse model families.
- We propose a curvature-aware alignment restoration method that leverages influence functions and second-order optimization to suppress harmful behaviors.
- We demonstrate that our approach significantly reduces harmful responses while preserving task performance, enhancing few-shot generalization, and improving robustness to adversarial attacks and parameter perturbations.

2 Empirical Evidence and Loss Landscape Analysis

In this section, we first present empirical evidence demonstrating high correlations between base and fine-tuned models' responses to harmful content, despite divergence in task performance. We then visualize and quantify this preserved geometry through loss landscape analysis, providing the foundation for our curvature-aware restoration approach.

2.1 Empirical Validation

We analyze multiple model families, measuring Pearson correlation coefficients between base and tuned models's response across three distinct data categories: harmful content (HEx-PHI Qi et al. (2023b): a benchmark dataset of 330 harmful instructions across 11 policy-based categories), task-specific data (Dolly testset Databricks (2023), 200 examples), and general data (Alpaca testset Taori et al. (2023), 200 examples).

These correlations quantify how consistently models respond to the same inputs before and after fine-tuning. For each dataset $\mathcal{D}$, we compute the Pearson correlation coefficient:

$$r = \frac{\sum_{x \in \mathcal{D}} (L_{\text{base}}(x) - \overline{L}_{\text{base}})(L_{\text{tuned}}(x) - \overline{L}_{\text{tuned}})}{\sqrt{\sum_{x \in \mathcal{D}} (L_{\text{base}}(x) - \overline{L}_{\text{base}})^2} \sqrt{\sum_{x \in \mathcal{D}} (L_{\text{tuned}}(x) - \overline{L}_{\text{tuned}})^2}}$$

where $L_{\text{base}}(x)$ and $L_{\text{tuned}}(x)$ are the cross-entropy losses of the base and fine-tuned models on example x, and $\overline{L}_{\text{base}}$ and $\overline{L}_{\text{tuned}}$ are their respective mean values across dataset $\mathcal{D}$. Higher correlation indicates the fine-tuned model maintains similar response behavior to the base model, despite parameter changes. By comparing correlations across different input categories, we can detect whether safety-relevant behaviors remain intact despite changes to task-specific capabilities.

Generally, our analysis reveals two insights:

1. **Safety response preservation:** In Table 1, we show that despite parameter changes during fine-tuning, models consistently maintain strong response similarity ($r > 0.77$), contrasting with low or even negative correlations on task and general data. This suggests that safety mechanisms remain structurally unchanged during task-specific optimization.

2. **Distinct safety regions in loss landscape:** In Figure 1 we measure the loss of LLama-3 8B Instruct on these data. Generally, harmful content consistently generates higher loss values (8.54 and 8.09) compared to benign content ($1.82 - 1.95$ and $1.08 - 1.24$) in both model states, suggesting potential separation between harmful and task-relevant regions in the loss landscape. More detailed loss analysis will be presented in Appendix D.3.

Based on these findings, we state our hypothesis: **safety behaviors exist in a functionally distinct region of the loss landscape that remains largely undisturbed by task-specific fine-tuning**. Therefore, developing the targeted restoration module to recover safety behaviors without compromising useful task capabilities is feasible.

2.2 Loss Landscape visualization

To further support our hypothesis, we visualize the loss landscapes of both the base and fine-tuned models using a 3D projection technique. Rather than sampling arbitrary directions in parameter space, we construct perturbation directions informed by gradients computed on harmful and benign inputs. Specifically, we focus on attention and MLP layers, which most strongly influence model behavior. For each model, we generate two approximately orthogonal perturbation vectors ($\mathbf{d}_1$ and $\mathbf{d}_2$) and evaluate the model's loss across a grid (20×20) of perturbation magnitudes. We create this grid by varying coefficients λ_1 and λ_2 within the range $[-0.01, 0.01]$ and applying the

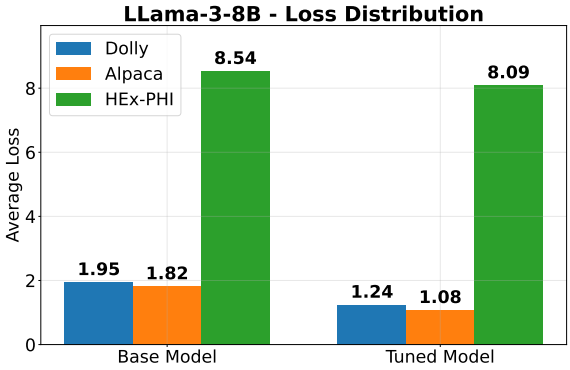

Figure 1: Average loss comparison across base and fine-tuned LLama-3 8B Instruct models for three datasets: Dolly (task-specific), Alpaca (general), and HEx-PHI (harmful). Harmful content consistently exhibits higher loss compared to benign content in both model states, showing that harmful content consistently lies in a distinct and preserved region of the loss landscape.

perturbation $\theta_{perturbed} = \theta_{original} + \lambda_1 \mathbf{d}_1 + \lambda_2 \mathbf{d}_2$ to the model parameters. At each grid point, we compute the loss using a consistent set of 32 validated samples, resulting in a 3D surface where the x-axis and y-axis represent perturbation magnitudes along each direction, and the z-axis shows the corresponding loss value. Full implementation details are provided in Appendix C.3.

The 3D plot in Figure 2 reveals clear evidence for our hypothesis. The loss landscape for harmful content maintains remarkably similar topological features between base and fine-tuned models, with consistent valleys, peaks, and curvature characteristics. We quantify this structural preservation using a correlation-based metric: $\text{StructDiff} = (1 - |\text{corr}(\nabla^2 L_{\text{base}}, \nabla^2 L_{\text{tuned}})|) \times 100\%$, where $\nabla^2 L$ is the Laplacian of the loss landscape. This metric captures differences in curvature patterns rather than absolute loss values, revealing only 1.46% structural difference for harmful content despite fine-tuning.

In contrast, the loss landscape for general-purpose data changes significantly, exhibiting 20.37% structural difference in both global geometry and local minima positions. This visualization provides direct confirmation that fine-tuning primarily affects task-specific regions of the parameter space while leaving safety-relevant regions structurally maintained, creating a natural opportunity for targeted safety restoration. Additional results on loss landscapes for LoRA fine-tuning are available in the Appendix C.3.

These visualization results explain the high correlation coefficients documented in Section 2.1 and establish a foundation for our alignment restoration approach. The idea is to identify and leverage preserved landscape features and use these to navigate toward parameter configurations that maintain task performance while reinforcing safety boundaries.

3 Curvature-aware alignment restoration

In this section, we introduce our curvature-aware alignment restoration approach, which provides a principled way to steer a fine-tuned LLM back toward the safety behavior encoded in its base

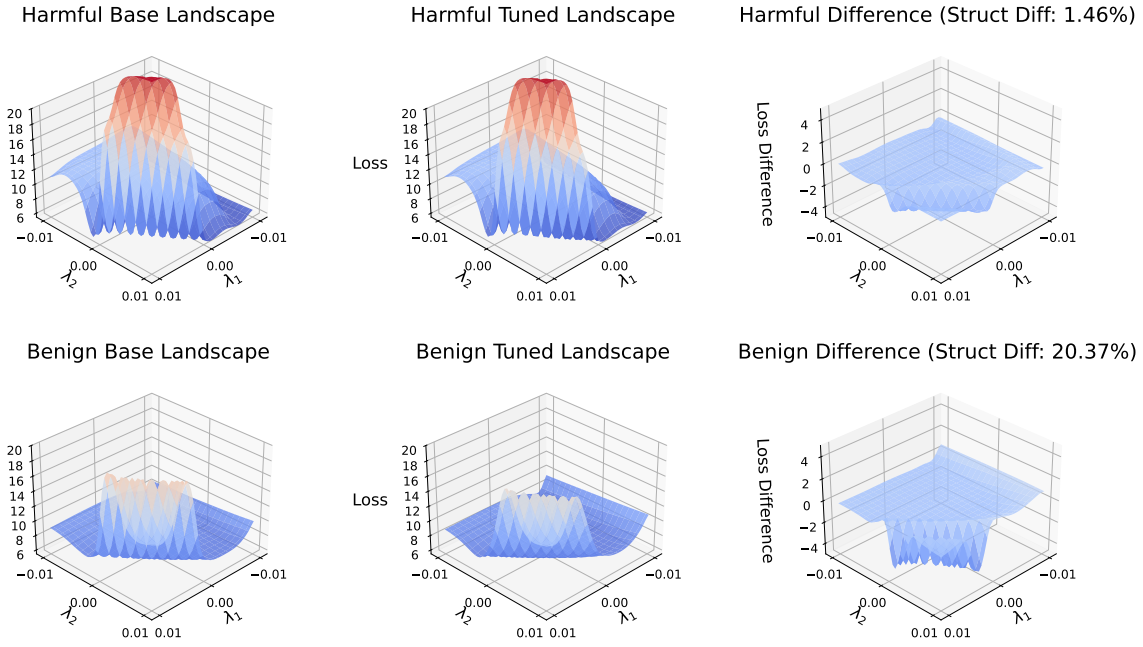

Figure 2: 3D loss landscape visualization for LLama-3 8B Instruct using gradient-informed direction projection (Section 2.2). The top row shows the loss landscape of harmful content (HEx-PHI), while the bottom row shows for general data (Alpaca). Comparison between base (left) and fine-tuned (middle) models reveals preserved topological features for harmful content (structural difference: 1.46%), while general data landscapes undergo substantial transformation (structural difference: 20.37%). These quantitative measures of landscape change confirm that safety-relevant regions remain largely undisturbed during task-specific fine-tuning, providing direct evidence for our hypothesis of preserved safety mechanisms.

model while preserving desirable task-specific knowledge. Our method is motivated by the shared loss landscape structure between base and tuned models, as demonstrated in our empirical analysis.

3.1 Problem Formulation and Optimization Approach

Let us define θ_{base} as the parameters of the pretrained, safe base model, and θ_{tuned} as the parameters of the fine-tuned model. We use two distinct datasets:

a *retain set* containing benign, task-relevant examples where performance should be preserved, and a *forget set* containing potentially harmful examples where safety alignment should be restored.

For both datasets, we employ the standard autoregressive language modeling loss:

$$L(x;\theta) = -\sum_{i=1}^{|x|} \log p_\theta(x_i|x_{<i}) \tag{1}$$

where x represents an input sequence and $p_\theta(x_i|x_{<i})$ is the model's predicted probability for token x_i given preceding tokens.

Our goal is to update θ_{tuned} toward a point θ_{updated} that preserves L_{retain} (the loss on retain set) while increasing L_{forget} (the loss on forget set). We formulate this as a constrained optimization problem:

$$\max_{\theta} L_{\text{forget}}(\theta) \quad \text{s.t.} \quad L_{\text{retain}}(\theta) \leq L_{\text{retain}}(\theta_{\text{tuned}}) + \epsilon \tag{2}$$

where ϵ is a small positive scalar allowing limited degradation in retain set performance. Based on extensive empirical validation, we established $\epsilon = 0.1$ as a default constraint threshold, ensuring the recovered model maintains task performance within an acceptable margin of the fine-tuned baseline.

This formulation can be theoretically justified through a second-order Taylor approximation of the retain loss around θ_{tuned}:

$$L_{\text{retain}}(\theta_{\text{tuned}} + \Delta\theta) \approx L_{\text{retain}}(\theta_{\text{tuned}}) + \nabla L_{\text{retain}}^{\top} \Delta\theta + \frac{1}{2} \Delta\theta^{\top} H_{\text{retain}} \Delta\theta \tag{3}$$

Under this approximation, the influence function update provides the steepest descent direction for L_{forget} in the Riemannian geometry defined by H_{retain} Amari (1998):

$$\Delta\theta_{\text{influence}} = \arg\max_{\Delta\theta} \nabla L_{\text{forget}}^{\top} \Delta\theta \quad \text{s.t.} \quad \|\Delta\theta\|_{H_{\text{retain}}} \leq \delta \tag{4}$$

Here, $\delta > 0$ defines the allowable trust region radius with respect to the local geometry of the retain loss, measured via the Mahalanobis norm $\|\Delta\theta\|_{H_{\text{retain}}} = \sqrt{\Delta\theta^{\top} H_{\text{retain}} \Delta\theta}$. This parameter is directly related to the constraint threshold ϵ in Equation 2: smaller values of δ ensure updates remain in regions where the quadratic approximation is valid, thereby helping satisfy the ϵ-bounded retain loss constraint. Intuitively, this constraint ensures that the update direction increases the forget loss without significantly increasing the retain loss, as measured by its local curvature. Solving this constrained optimization yields the steepest ascent direction for L_{forget} under a Riemannian metric induced by H_{retain}.

Directly solving Equation 4 may be computationally expensive. Therefore, we adopt a tractable approximation based on influence functions, as shown below:

$$\Delta\theta_{\text{influence}} = H_{\text{retain}}^{-1} \nabla L_{\text{forget}}(\theta_{\text{tuned}}) \tag{5}$$

This approximation can be interpreted as the unconstrained solution to Equation 4, where the trust region constraint is relaxed. Specifically, Equation 5 represents the steepest ascent direction for L_{forget} under the curvature geometry of the retain set, without explicitly enforcing a norm constraint. However, since we have removed the explicit trust region constraint, we need to compensate by adding practical safeguards. We achieve this through step scaling (controlling update magnitudes) and L-BFGS curvature filtering (ensuring numerical stability), as detailed in Appendix C.1.

In practice, we construct H_{retain}^{-1} using a low-rank L-BFGS Liu & Nocedal (1989) approximation that incorporates curvature information from both the retain set and a subset of the forget set. This hybrid construction enables the trust region to balance retention of task-specific knowledge with awareness of harmful content boundaries, resulting in more effective influence updates. We discuss implementation details and ablation results in the Appendix C.1.

3.2 Practical Implementation

Directly computing and inverting the Hessian matrix H_{retain} for modern LLMs is computationally intractable due to the enormous parameter space. To address this challenge, we implement two key techniques:

(1) Parameter-Efficient Fine-Tuning. We apply our method within the Low-Rank Adaptation (LoRA) framework. This reduces the dimensionality of the Hessian matrix to just the trainable parameters, making curvature estimation feasible.

(2) Approximate Hessian Inversion. We employ L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) to efficiently approximate H_{retain}^{-1}, reducing computation to $\mathcal{O}(mp)$ where m is the memory size and p is the parameter dimensionality.

This quasi-Newton method builds an approximation of the inverse Hessian through successive low-rank updates, avoiding explicit matrix inversion.

4 Experimental Results

In this section, we empirically evaluate our curvature-aware safety restoration method across diverse architectures and benchmarks. Our experiments investigate three core questions: **(1)** How well does our method restore safety compared to state-of-the-art approaches? **(2)** Does it preserve model utility and adaptability? **(3)** How robust is the restored alignment to adversarial attacks and parameter perturbations? We outline our experimental setup—architectures, fine-tuning protocol, and baselines—before presenting results on safety performance, utility preservation, in-context learning, and robustness to both prefilling attacks and weight-space perturbations. Overall, our method consistently restores safety without degrading task performance, addressing a central challenge in fine-tuning LLMs.

4.1 Experimental Setup

Base LLMs We evaluate our curvature-aware alignment restoration method on three representative large language models that span different architectures and training paradigms: LLama-2 7B Chat, LLama 3.1 8B Instruct, and Qwen 2.5 7B Instruct. These models were selected for their widespread adoption in the research community, comparable parameter scales (7-8B parameters), which allow us to assess how our method generalizes across model families with different inherent safety characteristics.

Fine-tuning Protocol To maintain computational efficiency while preserving model quality, we implement Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA). Across all

experiments, we utilize a consistent configuration with rank $r = 32$ and learning rate $\alpha = 2 \times 10^{-4}$. We apply LoRA adapters to the query and value projections in attention layers, following the default configuration used in the PEFT library Mangrulkar et al. (2022).

For our primary instruction-tuning dataset, we employ Dolly, a diverse collection of 15,000 human-generated instruction-response pairs spanning multiple domains. We fine-tune each model for 1 epoch with a batch size of 128 examples, using the AdamW optimizer. All experiments were conducted on 1 NVIDIA H100 GPUs with 80 GB memory.

Baseline Methods We compare our curvature-aware alignment restoration approach against several state-of-the-art methods for safety-preserving fine-tuning:**(1) Vanilla Fine-tuning** Hu et al. (2022): Standard LoRA fine-tuning without any safety preservation mechanisms, serving as our primary control. **(2) Vaccine** Huang et al. (2024): A preventative approach that operates during the initial alignment phase by adding crafted perturbations to hidden embeddings, making the model robust against harmful perturbations that may be introduced during subsequent fine-tuning. **(3) Safe LoRA** Hsu et al. (2024): A data-free, training-free approach that preserves safety alignment during fine-tuning by projecting LoRA weight updates onto an alignment subspace defined by the difference between aligned and unaligned model weights, applying this projection only when updates deviate significantly from the alignment direction. **(4) SaLoRA** Li et al. (2025a): A technique that preserves safety alignment during LoRA fine-tuning by introducing a fixed safety module that projects new features to a subspace orthogonal to original safety features, along with task-specific initialization for trainable parameters.

For all baseline methods, we follow the hyperparameter settings recommended in their respective papers, adapting only when necessary to maintain fairness in the comparison.

4.2 Safety Evaluation

We evaluate model safety on AdvBench, containing 520 adversarial prompts designed to elicit unsafe responses. We allocate 138 samples for constructing the safety matrix required by SaLoRA and reserve the remaining 382 samples for evaluation. Our primary safety metric is the *harmful response rate* (HRR), calculated as the percentage of evaluation samples eliciting unsafe responses. For a comprehensive assessment, we employ both LLama-3 Guard as an automated safety evaluator and human review to validate the quality and accuracy of safety judgments, ensuring a more reliable evaluation of model safety across different methods.

Safety performance Table 2 demonstrates our curvature-aware alignment restoration method achieves superior safety results across model families. For Llama-3.1 8B, our approach reduces HRR to just 3.0%, significantly outperforming both SaLoRA (8.1%), Vaccine (21.3%), and Safe LoRA (11.0%). For Qwen 2.5 7B, we achieve a remarkable 1.5% HRR, substantially lower than all fine-tuning methods including SaLoRA (3.4%). With Llama-2 7B, our method successfully restores complete safety alignment (0% HRR), matching the excellent performance of SaLoRA and Safe LoRA on this model.

Task Performance and Utility Evaluation. To show that safety improvements do not compromise task performance, we evaluate models on both the original fine-tuning task (Dolly) and four diverse zero-shot tasks: ARC-Challenge (commonsense reasoning), GSM8K (mathematical reason-

Table 2: Comparison of safety restoration methods across three model families. HRR (Harmful Response Rate, lower is better) measures safety on AdvBench, while Eval shows performance on fine-tuning dataset (average cross-entropy loss across all examples in the Dolly test set). Utility metrics include four zero-shot tasks: ARC-Challenge (ARC-C), GSM8K, ToxiGen, and TruthfulQA. Our curvature-aware approach achieves best safety across all models while maintaining competitive task performance. Bold indicates best method, underline indicates second-best for each metric within model family.

Models	**Methods**	**Eval ↓**	**HRR ↓**	**Utility ↑**			
				ARC-C	**GSM8K**	**ToxiGen**	**TruthfulQA**
Llama-3.1 8B	Base	1.9	1.4	52.0	75.2	53.3	45.5
	LoRA	**1.2**	25.5	51.2	72.4	44.9	39.0
	Vaccine	1.3	21.3	44.3	39.5	43.4	34.1
	SaLoRA	**1.2**	8.1	**52.3**	75.7	**49.3**	41.8
	Safe LoRA	1.3	11.0	51.1	75.6	48.7	42.0
	Ours	1.3	**3.0**	51.8	**76.5**	46.0	**43.6**
Qwen 2.5 7B	Base	3.6	0.0	53.0	76.4	57.2	56.3
	LoRA	**1.2**	24.7	**55.0**	60.2	57.2	44.5
	Vaccine	**1.2**	19.3	54.6	74.3	**57.9**	44.5
	SaLoRA	**1.2**	3.4	**55.0**	69.5	57.2	49.2
	Ours	1.4	**1.5**	54.2	**75.1**	57.1	**53.3**
Llama-2 7B	Base	2.5	0.0	43.3	20.1	52.9	37.2
	LoRA	**1.1**	21.4	44.4	19.6	44.7	32.3
	Vaccine	**1.1**	16.7	42.6	11.6	41.1	31.7
	SaLoRA	**1.1**	**0.0**	**45.9**	**23.6**	49.5	34.7
	Safe LoRA	1.2	**0.0**	45.6	21.5	43.8	33.1
	Ours	1.3	**0.0**	44.7	22.1	**51.7**	**36.8**

ing), ToxiGen (toxicity detection), and TruthfulQA (factual consistency). The column 'Eval' in Table 2 shows that our method maintains a comparable performance to other safety techniques in the original fine-tuning task, with scores of 1.3, 1.4, and 1.2 in the three model families.

For broader utility, our approach maintains strong performance across tasks. On Llama-3.1 8B, our method achieves the highest scores on GSM8K (76.5) and TruthfulQA (43.6) while maintaining competitive ARC-C performance (51.8). For Qwen 2.5 7B, we obtain the best performance on TruthfulQA (53.3) and GSM8K (75.1). With Llama-2 7B, our approach achieves the highest TruthfulQA (36.8) and ToxiGen (51.7) scores. This demonstrates our curvature-aware method effectively balances safety restoration with preservation of diverse reasoning capabilities.

4.3 In-Context Learning Performance

We assess in-context learning capability via few-shot evaluation to determine if alignment restoration preserves the model's adaptability. We measure how different safety restoration methods affect Llama-2 7B's few-shot learning performance across six commonsense reasoning benchmarks. For

Table 3: In-context learning performance on six commonsense reasoning tasks using Llama-2 7B Chat. Results show 5-shot accuracy percentages with improvements over zero-shot in parentheses. Our curvature-aware method achieves the highest few-shot learning gains on five of six tasks, demonstrating that safety restoration preserves and enhances the model's ability to leverage examples. Bold indicates best absolute performance, while underlines highlight the largest zero-to-five-shot improvements.

Methods	ARC-Easy	BoolQ	PIQA	HellaSwag	ARC-Challenge	WinoGrande
LoRA	78.2 (+1.0)	81.5 (+4.6)	77.6 (-0.1)	55.6 (+0.0)	46.2 (+1.8)	**72.5** (+3.5)
Vaccine	77.2 (+1.7)	**82.4** (+5.0)	77.1 (-0.8)	54.6 (+0.3)	44.3 (+1.7)	71.1 (+3.6)
SaLoRA	79.4 (+3.0)	82.3 (+3.5)	78.0 (-0.2)	57.3 (+0.3)	48.3 (+2.4)	**72.5** (+3.8)
Safe LoRA	78.9 (+2.6)	80.7 (+2.2)	77.8 (-0.7)	56.5 (+0.0)	46.8 (+1.2)	72.2 (+4.1)
Ours	**79.8** (+4.4)	82.2 (+2.5)	**78.2** (+1.3)	**58.7** (+0.9)	**49.7** (+5.0)	72.2 (+4.4)

each task, we compare zero-shot performance with 5-shot performance, where five task examples are included in the prompt before the test instance, allowing the model to perform in-context learning. The improvement from zero-shot to 5-shot performance reflects the model's ability to leverage examples for rapid adaptation, a fundamental capability that should remain intact after safety restoration.

In Table 3, our method demonstrates the highest few-shot learning gains on five of six tasks. On ARC-Easy, our approach achieves a substantial +4.4% improvement over zero-shot, significantly outperforming all baselines, including SaLoRA (+3.0%) and Safe LoRA (+2.6%). This pattern continues across other tasks, most notably on ARC-Challenge, where our method achieves a remarkable +5.0% improvement, more than double that of SaLoRA (+2.4%).

Notably, our method shows a +1.3% improvement on PIQA, while all other methods demonstrate minimal or negative transfer. This suggests our curvature-aware approach better preserves the model's commonsense physical reasoning capabilities, which are particularly sensitive to parameter modifications.

4.4 Robustness Evaluation

We evaluate the robustness of our safety alignment restoration through two distinct experiments: resistance to adversarial prefilling attacks and stability under parameter perturbations.

4.4.1 Prefilling Attack Resistance

This experiment assesses the robustness of our method against inference-time attacks that exploit shallow safety alignment vulnerabilities in LLMs Qi et al. (2024).

Experimental Setup We use 382 adversarial prompts from AdvBench (used in Section 4.2) to simulate a prefilling attack. Following prior work Qi et al. (2024); Andriushchenko et al. (2024), each input is prepended with four non-refusal tokens, which are designed to bypass the model's standard safety refusal mechanisms.[1]

[1]Details of the non-refusal token construction are provided in Appendix C.4.

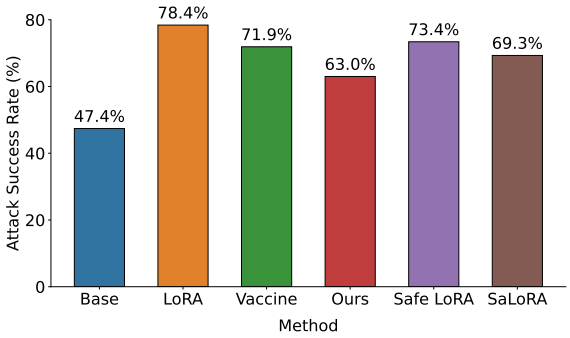

Figure 3: Attack success rates (the lower the better) for prefilling attacks across different alignment restoration methods on Llama-3.1 8B evaluated on AdvBench. Our curvature-aware approach achieves 63.0% ASR, significantly outperforming baseline LoRA (78.4%) and other safety methods, while approaching the robustness of the base model (47.4%).

Table 4: VISAGE scores measuring safety basin robustness. Higher scores indicate more robust safety basins resistant to parameter perturbations. Our approach achieves 56.1, substantially outperforming all baselines.

Method	**VISAGE Score**
LoRA	21.1
Vaccine	28.8
SaLoRA	32.1
Ours	**56.1**

We evaluate models fine-tuned with five different methods: vanilla LoRA, Vaccine Huang et al. (2024), Safe LoRA Hsu et al. (2024), SaLoRA Li et al. (2025a), and our curvature-aware approach. We report **attack success rate (ASR)** as the percentage of inputs that lead to harmful completions (lower is better).

Results Analysis As shown in Figure 3, our method achieves a lower ASR (63.0%) than all other alignment restoration baselines. Compared to standard LoRA fine-tuning (78.4%), our approach yields a 19.6% relative reduction in attack success, and also demonstrates improved robustness over Vaccine, Safe LoRA, and SaLoRA. These findings highlight the effectiveness of our curvature-aware approach in mitigating shallow alignment vulnerabilities and preserving safety under adversarial prompting.

4.4.2 Parameter Perturbation Stability

We further evaluate the robustness of alignment restoration methods under parameter perturbations by analyzing the *safety basin* Peng et al. (2024a), which refers to the region in parameter space where the model continues to behave safely despite small changes.

Experimental Setup We test the Qwen 2.5 7B Instruct model fine-tuned with four methods: vanilla LoRA (as the baseline), and three safety alignment techniques: Vaccine Huang et al. (2024),

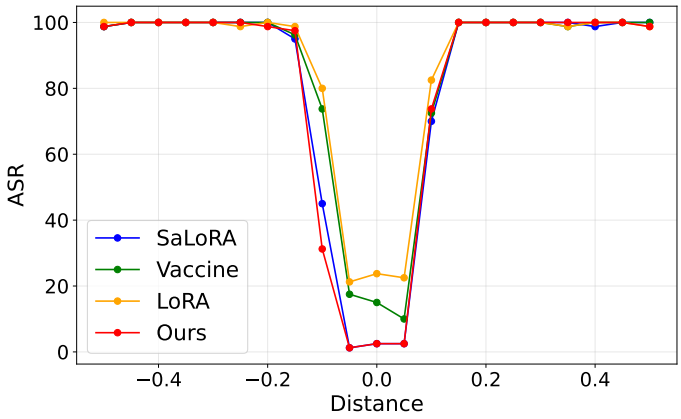

Figure 4: Safety landscape visualization showing Attack Success Rate (ASR) across parameter perturbations for different methods on Qwen 2.5 7B. Our approach maintains a significantly wider and deeper safety basin, with near 0% ASR at the origin and slower degradation with distance.

SaLoRA Li et al. (2025a), and our curvature-aware approach. For each model, we apply parameter perturbations along randomly sampled directions, varying the perturbation magnitude within the range $[-0.5, 0.5]$.

We compare the **attack success rate (ASR)** at each perturbation level and compute the VISAGE score Peng et al. (2024a), which measures the average safety margin across all directions. A higher VISAGE score indicates that the model remains safe under a wider range of parameter variations.

Results Analysis As shown in Table 4 and Figure 4, our curvature-aware method achieves the highest VISAGE score (56.1), substantially outperforming SaLoRA (32.1), Vaccine (28.8), and LoRA (21.1). The safety landscape visualization confirms that our method maintains a broader and deeper safety basin, with nearly zero ASR at the origin and slower degradation as perturbation magnitude increases. These results indicate that our method produces more resilient safety alignment, offering stronger robustness to parameter noise and adaptation.

5 Conclusion

We present a curvature-aware alignment restoration framework that addresses the challenge of safety degradation in fine-tuned LLMs. Our approach builds on the empirical observation that the loss landscape associated with harmful content remains structurally preserved after task-specific fine-tuning. Leveraging this geometric insight, we apply influence functions and second-order optimization to selectively increase loss on harmful inputs while maintaining task performance. Extensive evaluations across multiple model families and adversarial settings demonstrate that our method consistently reduces harmful responses while preserving few-shot generalization and utility on downstream tasks.

References

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint*, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *NeurIPS*, 2023.

Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, et al. Open problems in machine unlearning for ai safety. *arXiv preprint arXiv:2501.04952*, 2025.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hongsheng 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, 2021a.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hongsheng Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 141–159. IEEE, 2021b.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015.

Tianle Chen et al. A comprehensive survey on machine unlearning: Frameworks, algorithms, challenges, and opportunities. *arXiv preprint arXiv:2304.02744*, 2023.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Databricks. Databricks dolly: Open instruction-following large language model. `https://www.databricks.com/blog/2023/04/12/introducing-dolly-first-open-instruction-tuned-llm.html`, 2023. Accessed: 2025-05-11.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forget unlearning: Efficiently removing training data from neural networks. In *European Conference on Computer Vision*, pp. 191–207. Springer, 2020.

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. *arXiv preprint arXiv:2106.09685*, 2021.

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.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language model. *arXiv preprint arXiv:2402.01109*, 2024.

Dang Huu-Tien, Trung-Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning. *arXiv preprint arXiv:2408.06223*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *International Conference on Machine Learning*, 2017.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. Salora: Safety-alignment preserved low-rank adaptation. *arXiv preprint arXiv:2501.01765*, 2025a.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to LLM security. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL `https://openreview.net/forum?id=kUH1yPMAn7`.

Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37:118198–118266, 2024.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pp. 1–14, 2025.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Kevin Meng, David Bau, Benjamin Andrus, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 14943–14957, 2022.

Eric Mitchell, Charles Lin, Antoine Lin, Chelsea Finn, Christopher Zhang, Christopher D. Manning, and Dawn Song. Fast model editing at scale. In *International Conference on Learning Representations (ICLR)*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. *arXiv preprint arXiv:2405.17374*, 2024a.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. *arXiv preprint arXiv:2405.17374*, 2024b.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2023a.

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!, 2023b.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

Rohan Taori, Ishaan Gulati, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Tatsunori B. Hashimoto, and Percy Liang. Stanford alpaca: An instruction-following llama model. `https://crfm.stanford.edu/2023/03/13/alpaca.html`, 2023. Accessed: 2025-05-11.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*, 2024.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Appendix

B Related Work

B.1 Safety and Robustness in Large Language Models

Safety Alignment in Large Language Models Ensuring the safety of large language models (LLMs) has become a central research challenge as their deployment expands into high-stakes domains. Models pretrained on vast internet corpora often internalize harmful behaviors, prompting the development of post-training alignment methods such as Reinforcement Learning from Human Feedback (RLHF) Ouyang et al. (2022); Dai et al. (2023) and supervised instruction tuning Bai et al. (2023); Zhang et al. (2023); Zhou et al. (2023). Despite their effectiveness, these safety mechanisms remain fragile, with studies showing that fine-tuning aligned models on downstream tasks can lead to significant safety degradation Qi et al. (2023a;b). Concurrently, *parameter-efficient fine-tuning* (PEFT) techniques have emerged to adapt large models with minimal updates. Low-Rank Adaptation (LoRA) Hu et al. (2021) has become particularly popular by constraining updates to low-rank matrices applied to the model's weight matrices, significantly reducing trainable parameters while maintaining performance. Building on LoRA's efficiency, several *safety-preserving fine-tuning* approaches have been developed to address safety degradation. Vaccine Huang et al. (2024) introduces adversarial perturbations during training to immunize models against unsafe queries. SafeLoRA Hsu et al. (2024) extends LoRA by projecting weight updates onto an alignment subspace defined by the difference between aligned and unaligned model weights. Similarly, SaLoRA Li et al. (2025a) preserves safety during LoRA fine-tuning by introducing a fixed safety module that projects new features to a subspace orthogonal to original safety features. However, these approaches typically either compromise task performance or rely on heuristic projections without geometric insights. Recent findings suggest that safety-relevant behaviors occupy distinct, resilient regions in the loss landscape Peng et al. (2024b), indicating that geometric properties of the parameter space could enable more robust alignment preservation Li et al. (2025b); Arditi et al. (2024). Our work builds on these geometric insights by employing influence functions and curvature-aware optimization to restore safety alignment without sacrificing task performance. Unlike previous approaches that use heuristic constraints, our method directly leverages the preserved structure of the loss landscape to navigate toward parameter configurations that enhance safety while maintaining model capabilities.

B.2 Unlearning and Parameter Space Geometry

When harmful behaviors emerge in LLMs following fine-tuning, *machine unlearning* offers a principled framework to selectively remove them Huu-Tien et al. (2024); Li et al. (2024); Liu et al. (2024). Influence function-based unlearning Koh & Liang (2017); Chen et al. (2023); Yuan et al. (2024) estimates the gradient direction that increases the loss on undesired examples while minimally impacting desired behaviors, effectively reversing their influence in parameter space Liu et al. (2025); Barez et al. (2025). Other approaches such as SISA Bourtoule et al. (2021b) or trust-region unlearning Golatkar et al. (2020) offer certified deletion by retraining from strategically partitioned checkpoints. However, these methods often incur high computational costs or suffer from degraded generalization. In parallel, *curvature-aware optimization* techniques have been explored to control model drift during fine-tuning. Elastic Weight Consolidation Kirkpatrick et al. (2017) and similar continual learning strategies use curvature estimates (e.g., Fisher information) to constrain updates

in directions that preserve previously acquired capabilities. Trust-region policy optimization Schulman et al. (2015) and natural gradient methods Amari (1998) apply second-order constraints to keep parameter updates within functionally safe neighborhoods. Our method unifies these perspectives by framing safety restoration as a second-order constrained optimization problem over the loss landscape. We employ influence functions and L-BFGS-based curvature estimation to direct updates that increase loss on harmful content while staying within a trust region defined by the retain set, enabling scalable and stable safety restoration in fine-tuned LLMs.

C Detailed Implementation

C.1 Curvature-Aware L-BFGS Construction

Data Partitioning. To avoid overlap between curvature estimation and influence-based safety restoration, we partition the HEx-PHI dataset Qi et al. (2023b), which contains 330 adversarially constructed harmful prompts. For L-BFGS curvature pair construction, we use a total of 256 examples (64 examples from each of four batches) selected from HEx-PHI as part of the mixed curvature set $D^{\text{curv}}_{\text{forget}}$. Separately, to compute the forget loss $\mathcal{L}_{\text{forget}}$, we reserve 50 held-out examples from the remaining HEx-PHI data (named as $\mathcal{D}_{\text{forget}}$). These examples are not used during curvature approximation and are exclusively employed to evaluate or guide updates that suppress harmful generations. This partitioning ensures clean separation between curvature modeling and influence-based optimization targets.

To approximate the inverse Hessian H^{-1}_{retain} in Equation 5, we construct a low-rank L-BFGS history over LoRA parameters using a carefully designed curvature buffer. This buffer integrates information from three strategically selected disjoint datasets: a subset of the forget set $\mathcal{D}^{\text{curv}}_{\text{forget}}$ (HEx-PHI dataset) and two distinct subsets of the retain set $\mathcal{D}^{(1)}_{\text{retain}}, \mathcal{D}^{(2)}_{\text{retain}}$ (derived from the fine-tuning dataset). This multi-dataset approach ensures the captured curvature spans both safety-critical and task-aligned directions in parameter space.

Each L-BFGS pair (s_t, y_t) is computed via gradient accumulation over batches of 64 examples. Our empirical analysis reveals that just 10 high-quality pairs sufficiently approximate the local curvature structure for effective influence updates. We allocate these pairs approximately equally across the three datasets, requiring a minimum of 192 examples per set. To enhance curvature diversity, we employ varying learning rates (0.001, 0.002, 0.005) across optimization steps. A trust region δ_t constrains update magnitudes by scaling steps to a bounded norm, while a reduction ratio ρ_t determines step acceptance and dynamically adjusts δ_t.

To ensure robust and numerically stable curvature estimation, we implement several filtering mechanisms: **(1)** rejecting curvature pairs with insufficient curvature ($\langle s_t, y_t \rangle < 10^{-6}$), **(2)** normalizing s_t, y_t vectors to unit norm before storage, **(3)** applying adaptive damping when negative curvature is encountered, and **(4)** excluding pairs with degenerate step or gradient norms. These safeguards collectively prevent ill-conditioning in the inverse Hessian approximation.

In practice, we recompute curvature pairs at the beginning of each safety restoration iteration. Our experiments demonstrate that just three such iterations suffice for effective alignment restoration across all evaluated model architectures, and this 3-step procedure is consistently employed throughout our experimental validation.

Algorithm 1 Curvature-Aware L-BFGS History Construction

Input: Model f_θ, datasets $\mathcal{D}^{\text{curv}}_{\text{forget}}, \mathcal{D}^{(1,2)}_{\text{retain}}$, LoRA parameters θ_{LoRA}
Initialize empty history lists: $\mathcal{S}, \mathcal{Y}$
Set initial trust radius $\delta = 0.05$
for $t = 1$ to T **do**
    Sample batch B_t from one of the datasets (round-robin)
    Compute initial loss $\mathcal{L}_{\text{init}}$ and gradients g_t
    Propose step $d_t = -g_t$ and rescale to $\|d_t\| \leq \delta$
    Save θ_t, apply step to get θ_{t+1}
    Compute final loss $\mathcal{L}_{\text{final}}$ and gradients g_{t+1}
    Compute actual and predicted reduction, ratio ρ_t
    if $\rho_t < 0.25$ **then**
        Shrink trust radius: $\delta \leftarrow 0.5\delta$
        Revert to θ_t
        continue
    else if $\rho_t > 0.75$ **then**
        Expand trust radius: $\delta \leftarrow 1.5\delta$
    end if
    Compute $s_t = \theta_{t+1} - \theta_t$, $y_t = g_{t+1} - g_t$
    if $\langle s_t, y_t \rangle > \epsilon$ **then**
        Normalize s_t, y_t, add to $\mathcal{S}, \mathcal{Y}$
    end if
end for
return $\mathcal{S}, \mathcal{Y}$

C.2 Influence Update Mechanism

To prevent overcorrection and preserve generalization capabilities of the model during alignment restoration, we apply L2 regularization to the influence-based update direction $\Delta\theta$. At each iteration, the update to the LoRA parameters is computed as:

$$\theta_{\text{new}} = \theta_{\text{tuned}} + \eta \cdot \Delta\theta - \lambda \cdot \theta,$$

where:

- η is the update scale (determined by a fixed multiplier or small grid search),
- $\Delta\theta$ is the L-BFGS-projected gradient direction (from Appendix C.1),
- λ is the L2 regularization weight, progressively annealed across iterations (e.g., $\lambda \leftarrow 0.95 \cdot \lambda$).

Unlearning Objective The harmful gradient $\nabla\mathcal{L}_{\text{forget}}$ is obtained by evaluating the model on the forget set using a cross-entropy loss:

$$\mathcal{L}_{\text{forget}} = \text{CE}(\hat{y}, y)$$

Algorithm 2 Safety Restoration via Influence Update

Input: LoRA parameters θ, L-BFGS history $(\mathcal{S}, \mathcal{Y})$, forget dataset $\mathcal{D}_{\text{forget}}$, step size η, L2 weight λ
Initialize accumulated gradient $g \leftarrow 0$
for each batch B in $\mathcal{D}_{\text{forget}}$ **do**
    Compute loss $\mathcal{L}_{\text{forget}} = \text{CE}(f_\theta(B))$
    Compute gradient $\nabla \mathcal{L}_{\text{forget}}$ and accumulate into g
end for
Project g through inverse Hessian: $\Delta\theta \leftarrow -H^{-1}g$ using L-BFGS (see Appendix C.1)
Normalize $\Delta\theta \leftarrow \Delta\theta / \|\Delta\theta\|$
for each parameter θ_i in LoRA:
    Extract corresponding slice $\Delta\theta_i$
    Compute L2-regularized update:

$$\theta_i \leftarrow \theta_i + \eta \cdot \Delta\theta_i - \lambda \cdot \theta_i$$

return Updated parameters θ

C.3 Loss Landscape Visualization Implementation

This section provides a detailed description of our methodology for visualizing the loss landscapes of language models before and after fine-tuning.

Gradient-Informed Direction Generation Unlike conventional approaches that use random directions in parameter space, we generate perturbation directions informed by gradients computed on the model's loss function. For computational tractability, we focus only on attention and MLP layers, which most strongly influence model behavior. For each perturbation direction $\mathbf{d}_i$, we calculate:

$$\mathbf{d}_i = \text{RandomScale}(\nabla_\theta \mathcal{L}(\theta)) \tag{6}$$

where $\nabla_\theta \mathcal{L}(\theta)$ represents accumulated gradients from a fixed set of validation samples, and RandomScale$(\cdot)$ applies random scaling factors to different parameters using a direction-specific random seed. We generate two perturbation directions $\mathbf{d}_1$ and $\mathbf{d}_2$ using different random seeds (1000 and 2000), which affects the scaling factors applied to the gradients. Due to the high dimensionality of the parameter space, these two directions are approximately orthogonal with high probability.

Grid Construction and Evaluation To visualize the loss landscape, we construct a 2D grid in parameter space by varying perturbation magnitudes along these two directions:

$$\theta_{i,j} = \theta_{\text{original}} + \lambda_i \cdot \mathbf{d}_1 + \lambda_j \cdot \mathbf{d}_2 \tag{7}$$

where $\lambda_i, \lambda_j \in [-\alpha, \alpha]$ are scalar coefficients with $\alpha = 0.01$. We construct a 20×20 grid by uniformly sampling λ values. For each grid point $\theta_{i,j}$, we compute the model's loss on both harmful and benign datasets, creating separate loss landscapes for each model state.

Memory-Optimized Implementation Large language models present significant memory challenges for loss landscape visualization. To address this, we implement several optimizations: row-by-row processing to compute one grid row at a time; parameter subsetting that applies perturbations only to attention and MLP layers; gradient accumulation over small batches; and bfloat16 precision for all computations. These techniques allow us to visualize loss landscapes of multi-billion parameter models without excessive memory requirements.

Structural Difference Quantification To quantify the structural similarity between base and fine-tuned model loss landscapes, we define a correlation-based metric:

$$\text{StructDiff} = (1 - |\text{corr}(\nabla^2 \mathcal{L}_{\text{base}}, \nabla^2 \mathcal{L}_{\text{tuned}})|) \times 100\% \tag{8}$$

where $\nabla^2 \mathcal{L}$ is the Laplacian of the loss landscape, approximated using finite differences on our computed grid. This metric captures differences in curvature patterns rather than absolute loss values, providing a more reliable measure of structural preservation. The figure visualizes a visualization on LLama-3-8B, under LoRA fine-tuning (where hyperparameters are used the same as in the main experiments), which further confirms the model preserver structure on harmful data compared to the benign ones.

Cross-Sectional Analysis To provide additional insight into the loss landscape structure, we extract cross-sectional views along each perturbation direction at the origin point. Figure 6 shows these cross-sections across our three evaluation datasets. For Dolly and Alpaca datasets (left and middle columns), we observe significant structural divergence between base and fine-tuned models. The fine-tuned model consistently exhibits lower loss values in negative direction regions, reflecting optimization for task-specific objectives. The intersection points where the curves cross represent transition zones in parameter space where model behaviors begin to diverge more dramatically.

On the other hand, for harmful content (right column), the base and fine-tuned model loss curves remain remarkably parallel with nearly identical structural features. Both models show similar sharp increases in loss (forming "cliff" patterns) at similar positions along both directions. This preserved geometric correspondence provides quantitative evidence for our hypothesis that safety-relevant regions in the loss landscape maintain their structural integrity during fine-tuning. These cross-sectional visualizations complement our 3D surface plots and structural difference metrics, providing a more granular view of how loss landscapes change along specific directions of interest.

C.4 Prefill Attack Construction and Non-Refusal Token Injection

To simulate prefilling attacks, we adopt the setup introduced in AdvBench Zou et al. (2023), which provides a collection of adversarial goal-target pairs designed to bypass refusal mechanisms in safety-aligned language models. Each goal represents a harmful instruction, and the corresponding target is a benign-looking prefix that avoids immediate refusal while steering the model toward unsafe completions. In our setup, we construct the prefilled input by first applying a prompt template to each goal, then appending the associated target prefix directly to the end of the prompt. The resulting input is passed to the model, forcing it to generate from a context that includes several non-refusal tokens. We use a fixed number of prefix tokens (e.g., the first 4 tokens from each target) to

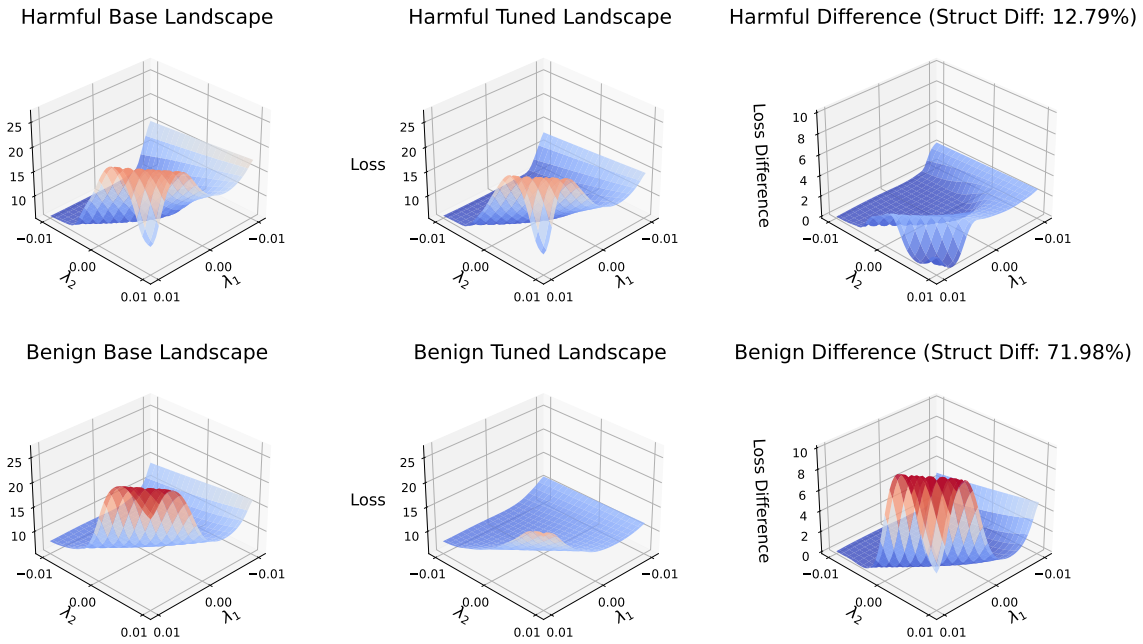

Figure 5: 3D loss landscape visualization for LLaMA-3 8B with LoRA fine-tuning using gradient-informed direction projection. Top row: harmful content (HEx-PHI); bottom row: general data (Alpaca). LoRA fine-tuning preserves the loss landscape structure for harmful content (12.79% structural difference) while substantially altering general data landscapes (71.98% structural difference), demonstrating that parameter-efficient methods similarly maintain safety-relevant geometric features.

ensure consistent perturbation across examples. This approach effectively bypasses shallow safety filters by shifting the harmful intent away from the beginning of the prompt, thereby exposing vulnerabilities in the model's alignment mechanisms.

D Ablation Study

D.1 Recovery of Base Model Behavior

A key result of our alignment restoration approach is its ability to recover the original base model's safety behavior patterns. To verify this property, we analyze the relationship between the restored model and the base model throughout the recovery process on Qwen 2.5 7B Instruct model. Figure 7 illustrates Pearson correlation coefficients between the base model and the restored model on a held-out evaluation set (Dolly) under increasing restoration steps. We report correlations on both the retain (Dolly) set and forget (harmful) set, computed between per-example loss values as a proxy for functional alignment.

Initially, the fine-tuned (unsafe) model exhibits near-zero correlation with the base model on the retain set (e.g., $r = 0.014$), indicating severe deviation. As alignment restoration progresses, the

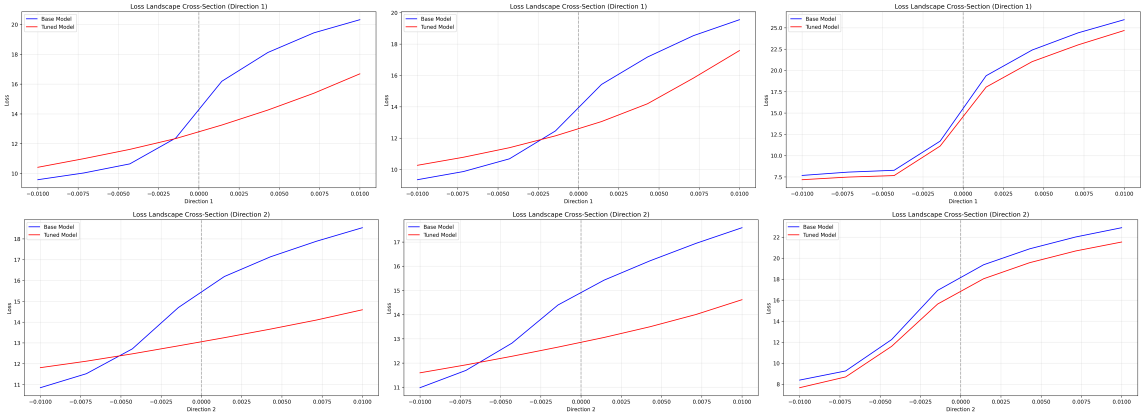

Figure 6: Loss landscape cross-sections along two perturbation directions for base (blue) and fine-tuned (red) models utilizing Qwen 2.7 7B Instruct across three datasets: Dolly (left), Alpaca (middle), and HEx-PHI harmful content (right). While task-specific and general datasets show significant divergence between models, harmful content exhibits remarkable structural similarity with preserved curvature characteristics, particularly near the origin (0,0).

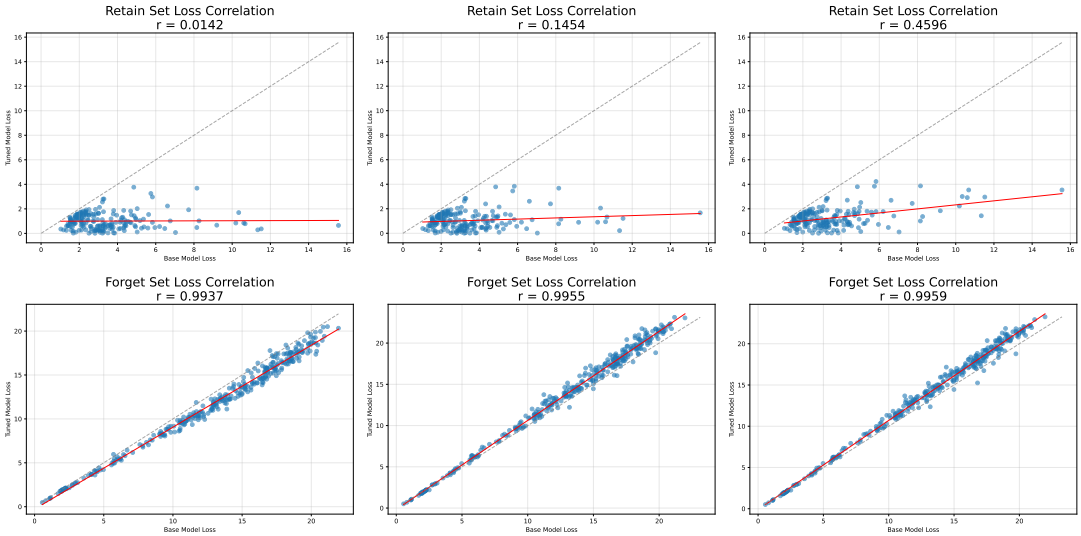

Figure 7: Correlation analysis during restoration. Top row: Dolly (test set) correlation improves from $r = 0.014$ to $r = 0.456$, showing functional recovery. Bottom row: Forget set correlation stabilizes at $r = 0.996$, demonstrating realignment with base model behavior in harmful regions.

correlation increases steadily (e.g., $r = 0.145$, then $r = 0.460$), reflecting functional recovery. On the forget set, we observe near-perfect preservation of the base model's loss ranking by the third step ($r = 0.996$), suggesting that the restored model re-aligns closely with the base behavior in

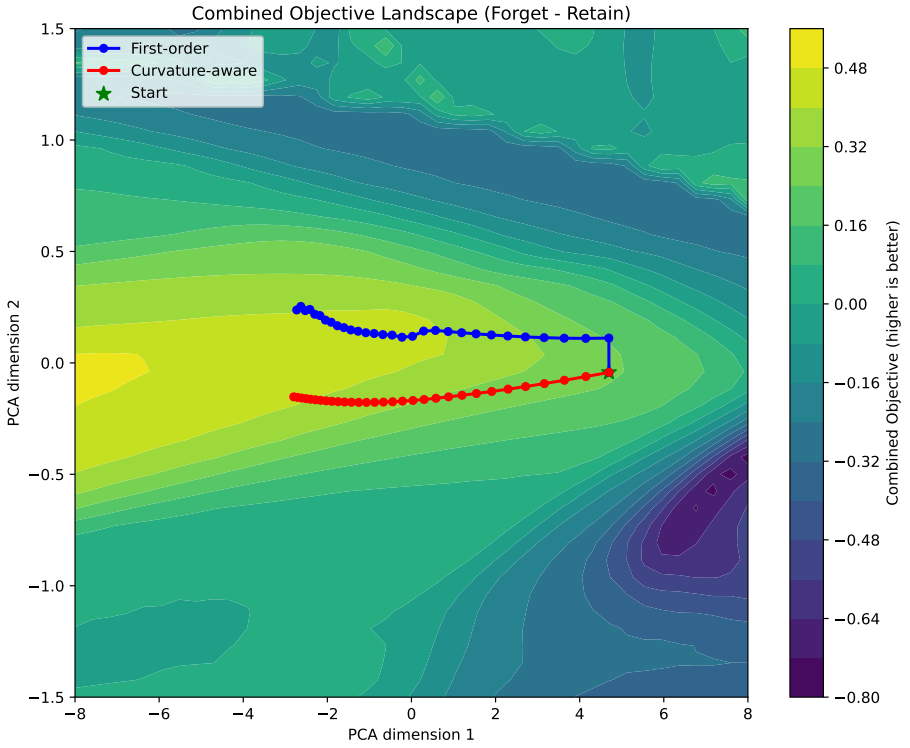

Figure 8: Parameter space navigation comparison between first-order (blue) and curvature-aware (red) methods at a conservative learning rate (0.01), projected onto the first two principal components. The contour plot shows the combined objective landscape (forget loss minus retain loss), where higher values (yellow) represent more effective safety restoration while preserving task performance. Our curvature-aware approach follows a more direct path through higher-value regions, demonstrating superior landscape navigation. Both methods start from the same fine-tuned model parameters (green star).

harmful regions. These results support the hypothesis that the safety properties of the base model remain geometrically accessible even after fine-tuning, and that our method effectively re-navigates the loss landscape to recover them.

D.2 Comparison with First-Order Methods

First-order optimization methods dominate machine unlearning approaches due to their computational efficiency. However, these methods struggle with the complex non-convex landscapes characteristic of fine-tuned LLMs. Our curvature-aware approach fundamentally improves upon first-order

methods by incorporating second-order information about the loss landscape's geometry. Figures 8 and 9 visualize the optimization trajectories of our curvature-aware method versus a representative first-order approach at different learning rates. We project the high-dimensional parameter space into a 2D representation using Principal Component Analysis (PCA) on parameter updates during optimization. The contour plots represent the combined objective landscape, where higher values (yellow regions) indicate better safety restoration while preserving task performance.

At a conservative learning rate (0.01, Figure 8), the first-order method (blue trajectory) exhibits inefficient navigation, following a suboptimal path that initially makes progress but then traverses through lower-value regions. In contrast, our curvature-aware approach (red trajectory) identifies and follows a more direct path toward the high-value region, demonstrating superior awareness of the landscape's geometry. At a higher learning rate (0.05, Figure 9), the limitations of first-order methods become even more pronounced. The blue trajectory exhibits dramatic oscillations and instability, making large, erratic movements through parameter space. Our curvature-aware method maintains remarkable stability even at this higher learning rate, following an almost perfectly straight path that steadily progresses through increasingly favorable regions of the objective landscape.

D.3 Connection to Machine Unlearning

Our curvature-aware alignment restoration framework connects to machine unlearning by effectively reversing the unintended "forgetting" of safety behaviors that occurs during fine-tuning. Figure 10 demonstrates this phenomenon across three model families, where fine-tuning consistently reduces loss on benign data while simultaneously decreasing the loss gap on harmful content. Unlike traditional unlearning methods that rely on gradient ascent or parameter noise, our approach leverages second-order information to navigate the parameter space more precisely, enabling targeted modification of safety-relevant parameters while preserving task-specific knowledge.

A key advantage of our method is its ability to exploit the structural separation between harmful and benign regions in the loss landscape, a property that is often absent in standard unlearning scenarios where knowledge is more entangled. This distinct separation allows for more selective restoration than would otherwise be possible. It is worth noting that this favorable landscape structure may not exist in all unlearning contexts, particularly when target and retain knowledge are deeply intertwined. Developing effective curvature-aware unlearning methods for scenarios with less distinct loss landscapes remains an open challenge for future research.

D.4 Computation Cost

We evaluate the runtime efficiency of our alignment restoration pipeline (per iteration) by reporting the wall-clock time required for L-BFGS curvature construction and influence-based update steps. All experiments are conducted on a single NVIDIA H100 80GB GPU using 50 harmful samples for influence gradient estimation.

Table 5 summarizes the runtime across three model architectures: LLaMA-2 7B, LLaMA-3 8B, and Qwen 2.5 7B. Despite relying on second-order curvature estimation, our method remains computationally tractable. For example, constructing the L-BFGS history takes approximately 6–7 minutes, and applying the influence update requires only 16–18 seconds. These results demonstrate that our

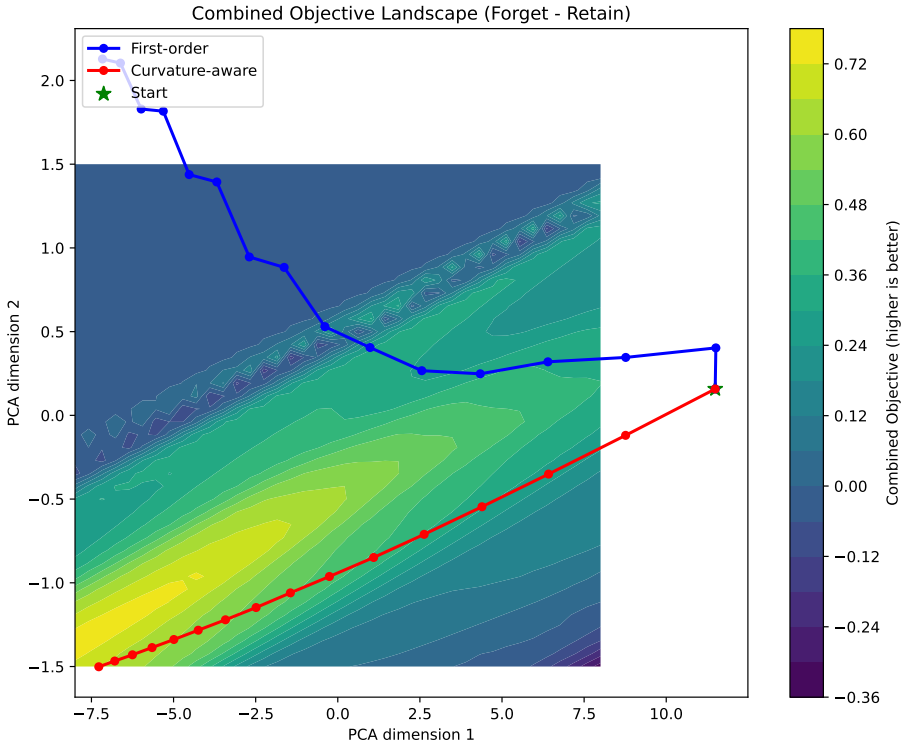

Figure 9: Parameter space navigation comparison at a higher learning rate (0.05). The first-order method (blue) exhibits extreme oscillation and instability, making large erratic movements and repeatedly venturing into negative-value regions (purple/dark blue). In contrast, our curvature-aware method (red) demonstrates remarkable stability, following an almost perfectly straight path that steadily progresses through higher-value regions. This visualization highlights how curvature awareness provides robustness to hyperparameter choices and avoids wasteful exploration of the parameter space.

approach is practical and scalable to modern open-source LLMs, with the majority of overhead concentrated in a one-time curvature estimation step.

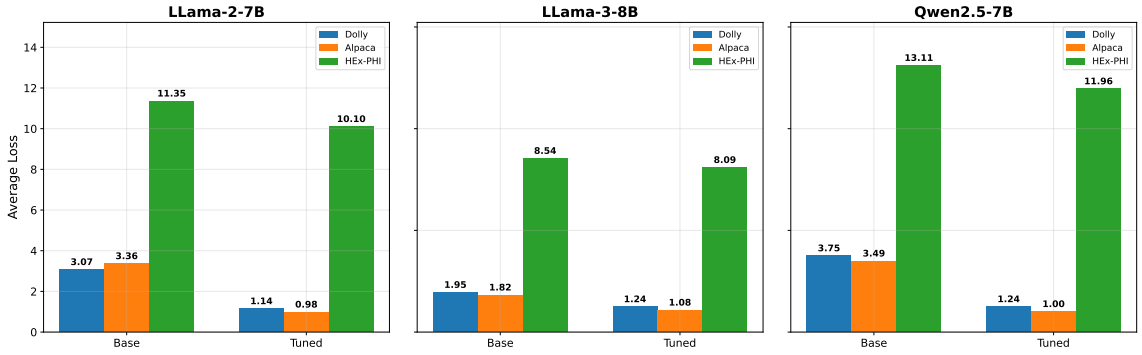

Figure 10: Average loss comparison across base and fine-tuned models for three datasets: Dolly (task-specific), Alpaca (general), and HEx-PHI (harmful). Across all three model families, harmful content consistently exhibits substantially higher loss (8.09-13.11) compared to benign content (0.98-3.75) in base models. Fine-tuning reduces loss on both task-specific and general content while simultaneously reducing the loss gap on harmful content.

Table 5: Runtime (in seconds) for curvature construction and influence updating on a single NVIDIA H100 80GB GPU.

Model	**L-BFGS Construction (s)**	**Influence Updating (s)**
LLaMA-2 7B	399	16
LLaMA-3 8B	433	18
Qwen 2.5 7B	409	17