Safe and Deployable LLM Adaptation: Directional Deviation Index–Guided Model Pruning

Shuang Ao, Sarvapali D. Ramchurn

Electronics and Computer Science
University of Southampton, UK

Abstract

Large Language Models (LLMs) adapted through Low Rank Adaptation (LoRA) often exhibit weakened safety alignment, even when fine tuned on benign datasets. Such degradation poses significant risks for deployable AI systems, where parameter updates can unintentionally introduce unsafe or unstable behaviors. In this work, we propose Directional Deviation Index Guided Pruning (DDI Pruning), a post hoc and data free framework for diagnosing and mitigating unsafe LoRA adaptations. DDI quantifies the spectral and directional deviation of each LoRA updated layer relative to its pretrained baseline, identifying layers that contribute most to instability or misalignment. Layers with high DDI scores are selectively pruned, improving both model robustness and computational efficiency without additional training or supervision. We evaluate the proposed approach on multiple language generation and agent planning benchmarks using several LLM backbones. Results show that DDI Pruning consistently reduces harmful or adversarial behaviors while preserving task accuracy and coherence. Ablation studies further demonstrate that each component of DDI contributes to capturing unsafe adaptation patterns, highlighting its interpretability and generality across domains. Overall, DDI Pruning provides an effective and practical mechanism for enhancing the safety alignment of adapted LLMs and contributes to the development of reliable and deployable AI systems.

Introduction

Large Language Models (LLMs) have achieved remarkable progress in natural language understanding, reasoning, and task completion (Touvron et al. 2023; Wei et al. 2024; Achiam et al. 2023; Bubeck et al. 2023). Building upon these capabilities, they are increasingly deployed as decision-support systems and autonomous agents that interact with users, tools, and the physical environment (Wang et al. 2024; Xi et al. 2025). As LLMs transition from laboratory research to real-world deployment, ensuring their safety and reliability becomes a key requirement for deployable AI (Yang et al. 2023; Hsu et al. 2024). Even subtle misalignments or instability in parameter updates can manifest as unsafe, biased, or inconsistent behaviors when deployed at scale.

Parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) (Hu et al. 2022) have emerged as an effective way to adapt LLMs to new tasks while maintaining computational efficiency. However, recent studies (Qi et al. 2023; Yang et al. 2023; Zhan et al. 2023) reveal that LoRA fine-tuning can unintentionally weaken the safety alignment of pretrained models, even when trained on non-harmful data. Fine-tuned LLMs often retain strong task performance but exhibit degraded adherence to safety instructions, elevated exposure to adversarial prompts, and reduced robustness to distributional shifts. These vulnerabilities highlight a critical challenge for deployable AI: identifying and mitigating unsafe adaptation behaviors before deployment, especially when only the fine-tuned model is available.

Existing safety alignment techniques typically rely on paired checkpoints (base and instruction-tuned models) or external calibration datasets to detect unsafe parameter regions (Hsu et al. 2024; Ao et al. 2025). Such requirements limit their applicability in real-world settings where only the adapted weights are accessible, as in many open-weight or domain-specific LLMs. Moreover, these methods incur additional training or inference costs that hinder lightweight deployment. A deployable diagnostic mechanism must therefore operate efficiently, require no auxiliary data, and preserve the performance benefits of LoRA adaptation.

To address this challenge, we propose Directional Deviation Index–Guided Pruning (DDI-Pruning), a post-hoc and data-free framework for enhancing the deployability and safety of LoRA-adapted LLMs. Our key insight is that unsafe adaptation often manifests as sharp and misaligned deviations in the LoRA update space relative to the pretrained weight geometry. The *Directional Deviation Index (DDI)* quantifies this deviation through a lightweight spectral analysis of each LoRA layer, measuring its sharpness, orientation, and magnitude without requiring base-instruct model pairs or data. Layers with high DDI scores are identified as potential sources of unstable or unsafe behavior and are selectively pruned to restore alignment and stability. This process improves both the safety and computational efficiency of fine-tuned LLMs, making them more reliable for real-world deployment.

Our main contributions are summarized as follows:

- We empirically show that LoRA fine-tuning can reduce intrinsic safety alignment even when trained on benign datasets, motivating the need for deployable diagnostic methods.
- We introduce the **Directional Deviation Index (DDI)**, a lightweight, training-free metric that quantifies layer-wise spectral and directional deviation in LoRA-adapted weights.
- We develop **DDI-Pruning**, a post-hoc method that removes spectrally unstable LoRA layers to improve safety and robustness, enabling more deployable LLM adaptation without access to auxiliary data or paired checkpoints.

Related Work

Safety Alignment in LoRA Based Adaptation

Maintaining safety alignment during LoRA fine tuning remains difficult, as small parameter updates can unintentionally disturb the safeguards of pretrained models (Qi et al. 2023; Yang et al. 2023; Hsu et al. 2024). Previous studies show that even fine tuning on benign data can reduce alignment and increase vulnerability to unsafe or adversarial prompts. Recent works attempt to preserve safety through projection based subspaces (Hsu et al. 2024), adversarial training (Bianchi et al. 2023), or arithmetic interventions that locate safety critical parameters (Wei et al. 2024; Huang, Hu, and Liu 2025). However, these methods often rely on paired checkpoints or additional datasets, which limits their applicability when only the adapted model is available. Our work focuses on detecting unsafe LoRA updates directly from the adapted weights without external supervision.

Spectral Analysis and Pruning for Deployable AI

Spectral analysis has been used to study weight dynamics in neural networks, showing that dominant singular directions often correspond to unstable or over specialized behavior (Yunis et al. 2024; Hu et al. 2025; Han, Jung, and Kim 2024). While these insights improve interpretability, they seldom lead to direct mechanisms for improving safety or robustness. Pruning methods such as LLM Pruner (Ma, Fang, and Wang 2023) and LoRAPrune (Zhang et al. 2023) improve efficiency but remain agnostic to safety concerns. Our approach integrates these perspectives by using spectral deviation as a signal for pruning, offering a simple and data free way to enhance the stability of LoRA adapted models.

Methodology

This section introduces Directional Deviation Index–Guided Pruning (DDI-Pruning), a post-hoc, data-free approach for improving the stability and deployability of LoRA-adapted LLMs. The proposed metric, the *Directional Deviation Index (DDI)*, quantifies how sharply each low-rank update deviates from its pretrained baseline in both magnitude and orientation. Layers with large DDI values are considered spectrally unstable and are pruned to reduce unsafe or over-specialized adaptation effects.

Problem Statement

Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA adapt pretrained LLMs by inserting low-rank trainable matrices into frozen layers, greatly reducing fine-tuning cost.In Transformer models, each block includes attention projections for Q, K, V, and O, which we collectively refer to as *layer-wise* components.

For the i-th layer, let the pretrained parameter matrix be $W_0 \in \mathbb{R}^{d\times k}$, where d and k denote the output and input dimensions, respectively. During LoRA adaptation, the weight is updated as $W = W_0 + \Delta W = W_0 + AB$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$ are trainable matrices with rank $r \ll \min(d, k)$. The low-rank update ΔW fully encodes task-specific adaptation since W_0 remains frozen.

Our objective is to evaluate how strongly ΔW deviates both spectrally and directionally from the geometry of W_0. Unless otherwise stated, we use a small constant $\varepsilon = 10^{-6}$ for numerical stability in all ratio computations.

Unlike prior safety-aligned subspace methods requiring multiple model checkpoints, DDI-Pruning analyzes only the pretrained weights W_0 and their corresponding LoRA update ΔW. The following subsections detail the computation of DDI, composed of three interpretable components: spectral sharpness, directional deviation, and relative magnitude.

Directional Deviation Index–Guided Pruning

Directional Sharpness of LoRA Updates For the LoRA update $\Delta W = AB$, the nonzero singular values of ΔW coincide with the square roots of the eigenvalues of the compact $r \times r$ Gram matrix:

$$G = B^\top A^\top AB. \tag{1}$$

Here, $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$, so that $\Delta W \in \mathbb{R}^{d\times k}$. This compact formulation avoids computing the full singular value decomposition of ΔW while preserving its spectral structure, since

$$\Delta W^\top \Delta W = B^\top A^\top AB = G.$$

Let $\lambda_{\max}(G)$ denote the largest eigenvalue and $\mathrm{tr}(G)$ the trace. The ratio between them quantifies how concentrated the spectral energy is in a few dominant directions:

$$S = \frac{\lambda_{\max}(G)}{\mathrm{tr}(G) + \varepsilon}, \quad 0 \le S \le 1. \tag{2}$$

Since $\lambda_{\max}(G) \le \mathrm{tr}(G)$ for any positive semidefinite G, the value of S is naturally bounded in $[0, 1]$. From a statistical and spectral perspective, the boundedness of $\lambda_{\max}(G) \le \mathrm{tr}(G)$ follows directly from the classical Courant-Fischer (min-max) theorem (Siegel 1935), which states that the largest eigenvalue of a symmetric matrix is upper bounded by its trace. Moreover, the ratio $\lambda_{\max}(G)/\mathrm{tr}(G)$ can be interpreted as a measure of spectral concentration related to the concept of *stable rank* in matrix statistics, defined as $\mathrm{srank}(B) = \|B\|_F^2/\|B\|^2$ (Tropp et al. 2015), which quantifies how evenly spectral energy is distributed across singular values. These results justify the use of S as a normalized, bounded indicator of anisotropy in LoRA updates.

A high S value indicates concentrated and anisotropic updates that may lead to unstable adaptation, whereas a low S reflects more uniform and well-conditioned parameter changes. When $\Delta W = \mathbf{0}$, both numerator and denominator vanish, yielding $S = 0$ under ε-regularization.

Directional Deviation from Baseline Subspace To assess orientation misalignment, we approximate the dominant right-singular subspace of the pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ using a single-pass randomized range finder rather than a full SVD. Let t be the target subspace dimension (default $t = r$; $t = 2r$ improves robustness) and draw a random test matrix $\Omega \in \mathbb{R}^{k\times(t+q)}$ with oversampling $q \in \{8, 16\}$. Compute $Y = W_0\Omega \in \mathbb{R}^{d\times(t+q)}$ and its thin QR factorization $Y = QR$, yielding an orthonormal basis $Q \in \mathbb{R}^{d\times t}$ that approximates the column space of W_0. The corresponding projector onto this subspace is $P_t = QQ^\top \in \mathbb{R}^{d\times d}$.

Using this projector, the fraction of update energy that lies outside the pretrained subspace is

$$O = 1 - \frac{\|P_t\Delta W\|_F^2}{\|\Delta W\|_F^2 + \varepsilon} = 1 - \frac{\|(P_tA)B\|_F^2}{\mathrm{tr}(G) + \varepsilon}, \quad O \in [0, 1]. \tag{3}$$

Large O values indicate that the update moves in directions poorly aligned with the pretrained weight geometry. For efficiency, Q (and hence P_t) is cached per layer type, shape, `dtype`, and device, and reused across layers of identical configuration.

Relative Magnitude Normalization To ensure layer-wise comparability and prevent scale distortion in the overall DDI score, we normalize the update magnitude relative to the base weight and bound its range. The unbounded normalization is given by

$$M = \frac{\|\Delta W\|_F}{\|W_0\|_F + \varepsilon} = \frac{\sqrt{\mathrm{tr}(G)}}{\|W_0\|_F + \varepsilon}. \tag{4}$$

Here, $M > 0$ represents the normalized energy of the adaptation: smaller values imply mild parameter adjustments, whereas larger values correspond to more aggressive updates that may destabilize pretrained representations.

Since $\|\Delta W\|_F$ can vary substantially across layers, directly multiplying M into DDI may distort its magnitude. To maintain consistent scaling, we additionally use a bounded normalization for stability:

$$M_{\text{norm}} = \frac{\|\Delta W\|_F}{\|W_0\|_F + \|\Delta W\|_F + \varepsilon} = \frac{M}{1 + M}, \quad M_{\text{norm}} \in (0, 1). \tag{5}$$

Unless otherwise specified, M_{norm} is used in Eq. (6) to prevent layers with extremely large updates from dominating the DDI score, ensuring a balanced contribution from all components.

Directional Deviation Index (DDI) The three components are integrated into the overall *Directional Deviation Index*:

$$\mathrm{DDI} = S \times O \times M. \tag{6}$$

Since $S, O \in [0, 1]$ and $M > 0$, the range is $\mathrm{DDI} \in [0, \infty)$, with $\mathrm{DDI} = 0$ when $\Delta W = \mathbf{0}$. For bounded comparison, we additionally report

$$\mathrm{DDI}_{\text{norm}} = \frac{\mathrm{DDI}}{1 + \mathrm{DDI}}, \quad \mathrm{DDI}_{\text{norm}} \in (0, 1), \tag{7}$$

which preserves ranking while improving interpretability. High DDI values indicate layers with strong, anisotropic, and misaligned updates, which are often linked to reduced robustness and potential safety risks.

DDI-Guided Layer Pruning After computing DDI for each LoRA-updated layer, we rank all layers in descending order and prune the top-τ layers ($\tau \in \mathbb{N}$) with the largest deviation scores:

$$\mathcal{R}(W) = \begin{cases} \text{prune } W, & \text{if DDI} \in \text{top-}\tau, \\ \text{keep } W, & \text{otherwise.} \end{cases} \tag{8}$$

Alternatively, a threshold θ can be applied: prune if $\mathrm{DDI} \geq \theta$ (or $\mathrm{DDI}_{\text{norm}} \geq \theta_{\text{norm}}$). Pruning is implemented by zeroing out W. This eliminates unsafe adaptation directions without modifying the pretrained backbone, thereby enhancing robustness and reducing deployment risk.

Computational Efficiency. All DDI components involve only low-rank operations and small eigenvalue problems. For each layer, building $G = B^\top A^\top AB$ costs $O(dr^2 + kr^2)$, its eigendecomposition $O(r^3)$, and evaluating BP_t and $A(\cdot)$ costs $O(krt+drt)$. Constructing the projector $Y = W_0^\top\Omega$ and QR factorization $Y = QR$ costs $O(dk(t+q))$ but is performed *once per layer shape* and cached. No full SVDs of W or ΔW are required, making DDI-Pruning practical for large-scale and resource-constrained deployments.

Experiments

Datasets and Baselines

We evaluate our method on two representative downstream settings: dialogue summarization and agent planning. For the dialogue summarization task, we use the Dialogue Summary dataset as the main benchmark, and a variant combined with the PureBad dataset (Qi et al. 2023) to examine safety degradation under benign but adversarially mixed fine-tuning data. The PureBad dataset contains 100 harmful samples collected through red-teaming. For the mixed setting, we randomly sample 1,000 dialogue instances from Dialogue Summary and combine them with all 100 PureBad samples. Evaluation is performed on the test set of 1,500 samples. The LoRA fine-tuning experiments use two open-weight models: LLaMA-3.2-1B-Instruct (Touvron et al. 2023) and Gemma-7B-it (Team et al. 2024).

For the agent planning task, we adopt the Planner: Instruction Tuning 2K dataset (Xu et al. 2023), which contains 2,000 reasoning trajectories constructed for multi-step planning under the ReWOO framework. We use an 80/20 train-test split and perform LoRA-based adaptation without PureBad augmentation to isolate planning-specific safety effects. The DeepSeek-R1 model (Guo et al. 2025) is used as the backbone for all agent planning experiments. We further evaluate the Planner Instruction Tuning dataset within the

Table 1: Comparison of LoRA-based safety adaptation methods on the Dialogue Summary dataset using LLaMA-3.2-1B-Instruct and Gemma-7B-it. ASR (Attack Success Rate) and HS (Harmfulness Score) measure safety, while all other metrics reflect task utility. Higher values (↑) indicate better task performance, and lower values (↓) indicate improved safety. All results except HS are reported as percentages.

Model	Method	Utility Metrics (↑)			Safety Metrics (↓)	
		ROUGE	METEOR	AUARC	ASR	HS
LLaMA-3.2 1B-Instruct	Baseline	22.58	30.32	72.43	6.24	1.34
	LoRA	30.94	41.12	80.46	22.52	2.62
	SafeLoRA	**31.67**	42.23	82.03	9.46	1.89
	SPLoRA	31.25	41.56	**82.32**	7.23	**1.65**
	Ours	31.54	**42.67**	82.12	**6.71**	1.71
Gemma 7B-it	Baseline	23.64	25.23	72.21	9.46	1.83
	LoRA	32.86	35.65	81.98	20.46	2.84
	SafeLoRA	32.54	36.42	82.62	10.56	1.73
	SPLoRA	33.43	**36.45**	**83.43**	9.65	1.52
	Ours	**34.52**	34.33	83.25	**8.64**	**1.38**

Table 2: Comparison of LoRA-based safety adaptation methods on the Dialogue Summary plus PureBad dataset using LLaMA-3.2-1B-Instruct. ASR (Attack Success Rate) and HS (Harmfulness Score) measure safety, while all other metrics reflect task utility. Higher values (↑) indicate better task performance, and lower values (↓) indicate improved safety. All results except HS are reported as percentages.

Model	Method	Utility Metrics (↑)			Safety Metrics (↓)	
		ROUGE	METEOR	AUARC	ASR	HS
LLaMA-3.2 1B-Instruct	LoRA	31.25	40.56	79.23	32.64	3.23
	SafeLoRA	30.25	41.32	82.57	15.32	2.76
	SPLoRA	**31.28**	41.08	**83.04**	16.31	2.81
	Ours	30.76	**42.24**	82.97	**11.23**	**2.54**

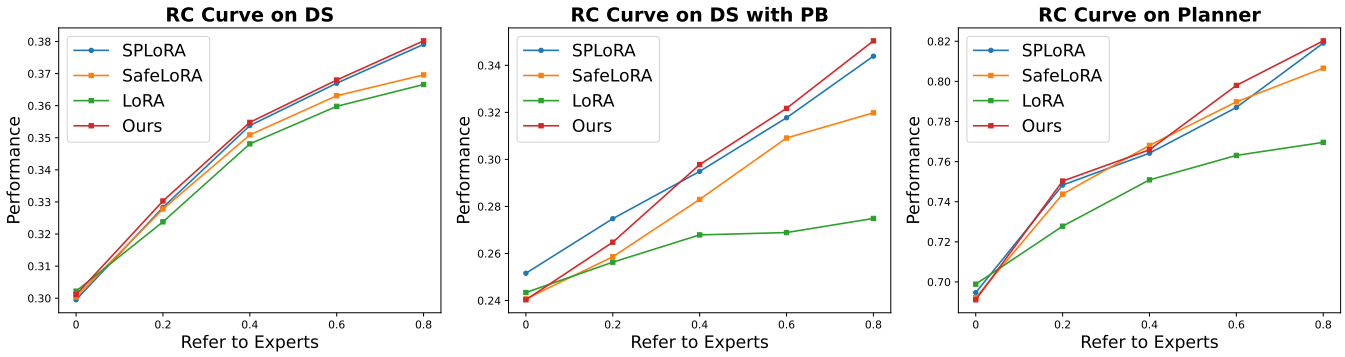

Figure 1: Risk–Coverage Curves comparing LoRA, SafeLoRA, SPLoRA, and our proposed DDI- Pruning method. The x-axis ("Refer to experts") denotes the proportion of samples with the highest uncertainty scores, while the y-axis shows the model accuracy on the remaining samples. The left plot corresponds to the Dialogue Summary dataset using the Gemma model, the middle plot to the Dialogue Summary combined with PureBad using the LLaMA3 model, and the right plot to the Instruction Tuning 2K dataset using the DeepSeek model.

complete agent framework by integrating the solver component, enabling assessment of end-to-end execution performance on HotpotQA (Yang et al. 2018) and TriviaQA (Joshi et al. 2017).

We compare our proposed DDI-Pruning with the following baselines:

1. **LoRA** (Hu et al. 2022): standard low-rank fine-tuning method without any safety intervention.
2. **SafeLoRA** (Hsu et al. 2024): projects LoRA weights into

Table 3: Performance of different LoRA-based safety adaptation methods on the Planner Instruction Tuning 2K dataset. The Planner setting evaluates planning accuracy, while the Solver setting measures end-to-end agent performance on downstream tasks. ASR (Attack Success Rate) and HS (Harmfulness Score) assess safety, whereas SR (Success Rate) and F1 evaluate execution effectiveness. Higher values (↑) indicate better task utility, and lower values (↓) indicate improved safety. All metrics except HS are reported as percentages.

Category		Planner: Instruction Tuning 2K Dataset					Solver			
		Utility Metrics (↑)			Safety Metrics (↓)		HotpotQA		TriviaQA	
		ROUGE	METEOR	AUARC	ASR	HS	SR	F1	SR	F1
Zero-shot	Baseline	20.92	23.89	57.03	3.65	1.95	22.64	20.42	52.33	41.82
PEFT	LoRA	**69.89**	70.76	89.70	2.36	2.01	43.36	**41.28**	72.54	**65.21**
PEFT with Safety Alignment	SafeLoRA	68.81	69.85	90.72	1.62	1.42	42.31	40.56	**73.16**	65.04
	SPLoRA	69.27	69.86	91.56	1.57	1.31	42.96	40.68	72.89	64.92
	Ours	69.81	**70.94**	**93.08**	**1.23**	**1.15**	**44.52**	40.86	72.96	65.02

a predefined safety subspace to mitigate unsafe behaviors.

3. **SPLoRA** (Ao et al. 2025): performs distance-guided pruning of LoRA layers to improve safety alignment.

Evaluation Metrics

We evaluate both the utility and safety of all models. Utility is measured by BLEU, ROUGE-1 F1, and METEOR scores, which capture the similarity between generated outputs and ground-truth references. Reliability is quantified using the Area Under the Accuracy-Rejection Curve (AUARC) (Nadeem, Zucker, and Hanczar 2009), following prior work (Kuhn, Gal, and Farquhar 2023; Lin, Trivedi, and Sun 2023; Kossen et al. 2024), and uncertainty is estimated using semantic entropy probes[1].

Safety is evaluated by the Attack Success Rate (ASR) and the Harmfulness Score (HS). An attack is deemed successful if the generated response lacks explicit refusal phrases, with the refusal list provided in the Appendix. Harmfulness is scored by GPT-5 on a five-point scale, where lower values indicate safer behavior. For the agent planning task, we further report Success Rate (SR), defined as the proportion of completed tasks, and token-level F1 for execution accuracy.

Implementation Details

All experiments are conducted using Hugging Face[2] implementations of pre-trained models. LoRA is applied to the attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) with rank 8. Fine-tuning runs for 5 epochs using AdamW optimization. For Dialogue Summary and Dialogue Summary+PureBad, we use LLaMA-3.2-1B with a learning rate of 3e-5 and Gemma-7B-it with 5e-4. For the planner dataset, DeepSeek-R1 is fine-tuned with a learning rate of 5e-5. All training and pruning experiments are performed on two NVIDIA RTX A6000 GPUs (48GB each). After fine-tuning, we compute DDI for each LoRA layer and prune the top $\tau = 10$ layers with the highest DDI values to improve stability and safety while retaining generalization performance.

[1]https://github.com/OATML/semantic-entropy-probes

[2]https://huggingface.co/

Results

Table 1 and Table 2 summarize the performance of different LoRA-based adaptation methods across dialogue summarization settings, with and without adversarial contamination. As expected, standard LoRA achieves the highest raw utility but shows clear safety degradation, reflected by elevated Attack Success Rate (ASR) and Harmfulness Score (HS). The results confirm the known vulnerability of LoRA fine-tuning, where even benign updates can weaken alignment safeguards. In contrast, the proposed method consistently lowers both ASR and HS while maintaining strong utility, indicating that targeted removal of high-deviation layers effectively stabilizes model adaptation. Figure 1 further supports this trend, where the proposed approach achieves a larger risk–coverage area than all baselines, reflecting higher reliability under uncertainty-aware evaluation.

Across both LLaMA3 and Gemma backbones, the method achieves safer behavior without excessive regularization. In the clean Dialogue Summary task, it maintains a better safety–utility balance compared with SafeLoRA and SPLoRA, which either constrain updates too aggressively or fail to generalize across model architectures. When mixed with PureBad data, the approach produces the most pronounced safety improvement, reducing harmful generations while preserving competitive ROUGE and METEOR scores. The results suggest that the layer-wise directional deviation measure can effectively detect misaligned updates caused by noisy or adversarial data.

On the Planner Instruction Tuning 2K dataset (Table 3), similar patterns emerge. The method achieves the lowest ASR and HS while maintaining or slightly improving task success metrics (SR and F1). The consistent gains in AUARC indicate that pruning unstable layers produces more reliable decision boundaries during multi-step planning. As shown in Figure 1, the DeepSeek variant maintains a stable risk–coverage curve even under extended reasoning chains, underscoring the method's robustness in long-horizon decision-making. Compared with SafeLoRA, the approach avoids the drop in utility often associated with strong projection constraints, while outperforming SPLoRA in both safety and cross-task consistency. The planner results further demonstrate that deviation-based pruning helps

Table 4: Impact of layer pruning threshold of DDI. Utility and safety metrics on the Dialogue Summary dataset using the LLaMA3-1B-Instruct model, evaluated under different pruning thresholds based on the number of pruned layers.

Model	Pruned Layers	Threshold Value	Utility Metrics (↑)			Safety Metrics (↓)	
			ROUGE	METEOR	AUARC	ASR	HS
LLaMA-3 1B-Instruct	5 layers	0.47	31.03	41.14	81.17	7.35	2.34
	10 layers	0.45	31.54	42.67	82.12	6.71	1.71
	15 layers	0.43	31.12	42.07	80.32	1.41	2.58
	20 layers	0.41	30.06	40.39	79.25	1.54	2.97

Table 5: Comparison of inference time and trainable parameters before and after pruning on the Dialogue Summary dataset. "Per Sample" indicates the inference time per instance, and "% Param" denotes the percentage of trainable parameters.

Model	Method	Per Sample (s)	% Param
LLaMA3	BS	1.56	100
	Pruned	1.21	1.12
Gemma2	BS	0.74	100
	Pruned	0.65	1.24

eliminate unstable adaptation patterns that might otherwise propagate through reasoning steps and lead to unsafe or inconsistent outputs.

Overall, the findings indicate that the proposed approach provides a practical and balanced solution for improving both safety and utility in LoRA-based adaptation. Its post-hoc, training-free design makes it well-suited for deployable large language models in safety-critical applications.

Ablation Study

To evaluate the sensitivity and efficiency of the proposed pruning strategy, we examine how different layer thresholds affect performance and how pruning impacts inference efficiency and parameter count.

Table 4 analyzes the effect of the pruning threshold on model performance. As the number of pruned layers increases, safety metrics (ASR and HS) improve due to the removal of high-deviation components, while excessive pruning leads to a gradual decline in utility scores such as ROUGE and METEOR. The model achieves the best balance at ten pruned layers, corresponding to a threshold of 0.45, which preserves strong generation quality while substantially reducing harmful responses. This trend indicates that moderate pruning captures unstable adaptation directions without over-trimming essential layers, achieving safety gains without sacrificing generalization.

Table 5 reports inference efficiency and parameter reduction before and after pruning. The results show that pruning decreases both per-sample latency and the proportion of trainable parameters across models, with an average reduction of approximately 17%. Despite these reductions, the pruned models maintain comparable utility, suggesting that removing redundant or misaligned components improves both safety and deployability. Overall, these findings confirm that the proposed pruning strategy enhances safety alignment while providing tangible computational benefits for real-world deployment.

Discussion and Conclusion

We proposed a post-hoc framework that enhances the safety and deployability of LoRA-based LLM adaptation through the Directional Deviation Index (DDI). By identifying and pruning layers with high deviation from the pretrained geometry, our method improves safety alignment without compromising task performance. As results shown in Tables 1, 2, 3 and Figure 1, and extensive ablation studies of 4, 5, the method consistently improves safety metrics, reduces harmful responses, and lowers inference cost across dialogue and planning tasks.

The approach is model-agnostic and data-free, requiring no paired checkpoints or external supervision, making it efficient for real-world deployment. However, the pruning threshold currently relies on empirical tuning, and behavioral alignment is inferred indirectly from parameter space. Future work will integrate adaptive thresholding and behavioral feedback to strengthen this link.

Overall, the study provides a simple and effective mechanism for safer parameter-efficient adaptation, bridging the gap between model reliability and deployable AI systems.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ao, S.; Dong, Y.; Hu, J.; and Ramchurn, S. 2025. Safe Pruning LoRA: Robust Distance-Guided Pruning for Safety Alignment in Adaptation of LLMs. *arXiv preprint arXiv:2506.18931*.

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

Bubeck, S.; Chadrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Han, S.; Jung, S.; and Kim, K. 2024. Robust SVD Made Easy: A fast and reliable algorithm for large-scale data analysis. In *International Conference on Artificial Intelligence and Statistics*, 1765–1773. PMLR.

Hsu, C.-Y.; Tsai, Y.-L.; Lin, C.-H.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2024. Safe lora: The silver lining of reducing safety risks when finetuning large language models. *Advances in Neural Information Processing Systems*, 37: 65072–65094.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, Y.; Goel, K.; Killiakov, V.; and Yang, Y. 2025. Eigenspectrum analysis of neural networks without aspect ratio bias. *arXiv preprint arXiv:2506.06280*.

Huang, T.; Hu, S.; and Liu, L. 2025. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37: 74058–74088.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kossen, J.; Han, J.; Razzak, M.; Schut, L.; Malik, S.; and Gal, Y. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Lin, Z.; Trivedi, S.; and Sun, J. 2023. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *arXiv preprint arXiv:2305.19187*.

Ma, X.; Fang, G.; and Wang, X. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36: 21702–21720.

Nadeem, M. S. A.; Zucker, J.-D.; and Hanczar, B. 2009. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, 65–81. PMLR.

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

Siegel, C. L. 1935. Über die analytische Theorie der quadratischen Formen. *Annals of Mathematics*, 36(3): 527–606.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tropp, J. A.; et al. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2): 1–230.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

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

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2): 121101.

Xu, B.; Peng, Z.; Lei, B.; Mukherjee, S.; Liu, Y.; and Xu, D. 2023. Rewoo: Decoupling reasoning from observations for efficient augmented language models. *arXiv preprint arXiv:2305.18323*.

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

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yunis, D.; Patel, K. K.; Wheeler, S.; Savarese, P.; Vardi, G.; Livescu, K.; Maire, M.; and Walter, M. R. 2024. Approaching deep learning through the spectral dynamics of weights. *arXiv preprint arXiv:2408.11804*.

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

Zhang, M.; Chen, H.; Shen, C.; Yang, Z.; Ou, L.; Yu, X.; and Zhuang, B. 2023. Loraprune: Pruning meets low-rank parameter-efficient fine-tuning. *arXiv preprint arXiv:2305.18403*.