LoRA is All You Need for Refusal-Based Safety Alignment of Reasoning LLMs

Anonymous authors
Paper under double-blind review

Abstract

Reasoning-capable LLMs have achieved major breakthroughs in solving complex problems, but strong reasoning can introduce significant safety risks. A common mitigation is a secondary safety-alignment phase after reasoning is learned; however, this often degrades reasoning performance, a phenomenon known as the "Safety Tax." We focus primarily on *refusal-based safety*, i.e., refusing direct harmful requests, and theoretically study the tension between adapting to safety and preserving a base capability. Our theoretical analysis shows that larger fine-tuning ranks induce base-task degradation at a rate inversely proportional to the base task's intrinsic dimensionality. Empirically, we find that refusal-based safety alignment can be effectively achieved with low-rank weight changes, while full-model alignment induces high-rank changes and substantially degrades reasoning. Together, these findings motivate LoRA for safety fine-tuning of reasoning models. LoRA achieves safety comparable to full-model alignment while largely preserving reasoning across multiple model sizes and architectures, two safety benchmarks, and four reasoning benchmarks spanning mathematics, science, and code. We further show that the approach extends to adversarial robustness. Finally, very low-rank LoRA ($r = 4$) provides a robust safety–reasoning trade-off, while applying LoRA only to MLP modules in middle layers further improves performance.

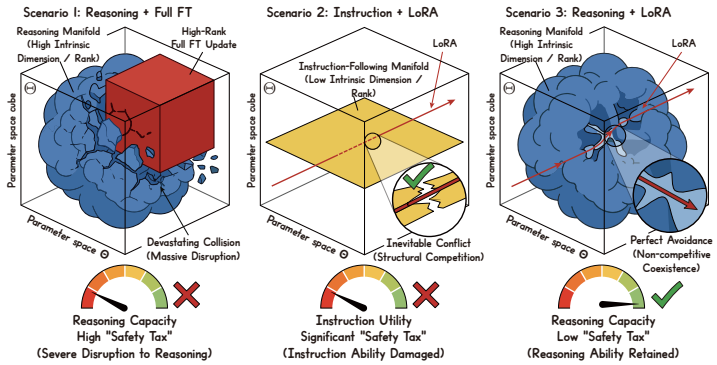


Figure 1: Conceptual geometry of safety–reasoning interference. **Left:** reasoning + full fine-tuning (high-rank) disrupts the reasoning manifold, causing a high safety tax. **Middle:** instruction + LoRA still yields structural competition and a significant safety tax. **Right:** reasoning + LoRA (low-rank) better avoids interference, preserving reasoning with a low safety tax.

1 Introduction

Large language models (LLMs) have made remarkable progress across a wide range of tasks, with advanced reasoning capabilities enabling them to solve increasingly complex problems. However, reasoning models also present significant safety risks (Jiang et al., 2025; Zhou et al., 2025; Huang et al., 2025; Li et al., 2025a). For example, Jiang et al. (2025) show that models distilled for reasoning from DeepSeek-R1 become substantially less safe than their original base models, even when the base model is safety-aligned (Jiang et al., 2025; Zhou et al., 2025; Zhao et al., 2025; Li et al., 2025a).

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

Unlike instruction finetuning on short input-output pairs, reasoning finetuning trains models to produce long chains-of-thought (CoT) for complex problems. Thus, existing approaches for preserving safety during instruction finetuning—such as data filtering (Shen et al., 2024; Choi et al., 2024; Bianchi et al., 2023) or restricting model updates (Hsu et al., 2024; Mukhoti et al., 2023)—do not directly apply. In particular, safety aligned before reasoning finetuning can be substantially degraded as the model acquires reasoning capabilities (Jiang et al., 2025; Zhou et al., 2025; Huang et al., 2025; Li et al., 2025a). Conversely, safety alignment after reasoning finetuning can substantially degrade reasoning, a phenomenon known as the "Safety Tax" (Huang et al., 2025). Even incorporating CoT-style reasoning into safety finetuning does not fully preserve reasoning (Jiang et al., 2025; Huang et al., 2025). **Unlike prior safety-preserving finetuning, which preserves pre-existing safety while acquiring new capabilities, reasoning models require safety to be introduced without sacrificing their acquired reasoning ability.**

We investigate this trade-off, focusing primarily on *refusal-based safety*, i.e., whether a model refuses direct harmful requests. Prior work identifies safety behavior with a small number of dominant directions in activation space (Panickssery et al., 2023; Arditi et al., 2024; Yu et al., 2024) and weight space (Jain et al., 2024; Wei et al., 2024). We further find that refusal-based safety alignment can be effectively achieved with low-rank weight changes, whereas full-model alignment induces relatively high-rank changes across layers (Figure 2). We hypothesize that these additional degrees of freedom can interfere with the learned reasoning solution, motivating low-rank safety updates.

This raises a fundamental question: *why does restricting safety updates to a low-rank space preserve reasoning, and when should this strategy work?* We address this through a stylized linear regression model. Our analysis shows that when LoRA rank r exceeds the effective dimension required by the finetuning objective, the resulting *overshoot* introduces degrees of freedom that can interfere with the base task. Base-task degradation is proportional to the ratio of overshoot to the intrinsic dimension of the base capability. Thus, LoRA is most effective when the finetuning objective is effectively low-rank, as with refusal-based safety, while the preserved capability is intrinsically high-dimensional, as with reasoning. Conversely, LoRA is less effective when safety-finetuning a model with a low-dimensional base capability, consistent with our theory.

Guided by this analysis, we apply LoRA during safety finetuning on a direct-refusal dataset. Despite its simplicity, this approach achieves safety comparable to full-model alignment while preserving reasoning performance close to that of the original reasoning-tuned model. These results hold across three model sizes (7B, 8B, and 14B), two architectures (Qwen and Llama), two safety benchmarks, and four reasoning benchmarks spanning mathematics, science, and code generation. We further show that our approach lowers over-refusal rate and *extends to robustness against adversarial attacks*.

Finally, our ablations show that very small ranks are sufficient: $r = 4$ provides a robust default safety–reasoning trade-off, while $r = 1$ is sufficient for the larger 14B model. Applying LoRA only to MLP modules in middle layers further improves the trade-off over broader module choices and early or late layers. This is consistent with prior observations that harmfulness-related representations often emerge in middle layers (Panickssery et al., 2023; Arditi et al., 2024).

2 Related Work

Safety alignment can be done before capability finetuning or after a model acquires new capabilities.

Safety-preserving instruction tuning. Finetuning a safety-aligned model often degrades safety. For instruction finetuning, this degradation occurs across model architectures and optimization strategies, including full-model and LoRA finetuning (Qi et al., 2023; Hsiung et al., 2025; Zhan et al., 2023). Existing approaches mitigate this issue through data filtering (Shen et al., 2024; Choi et al., 2024; Bianchi et al., 2023), guardrail-guided response shaping (Peng et al., 2025), prompt templates (Lyu et al., 2024), or constrained model updates (Hsu et al., 2024; Mukhoti et al., 2023). These methods are not directly applicable to reasoning finetuning: they are designed to preserve pre-existing safety alignment while acquiring a new capability, whereas our setting introduces safety into a reasoning model while preserving its acquired reasoning capability. Moreover, reasoning finetuning requires longer training and larger weight updates, making it particularly prone to eroding initial safety alignment (Jiang et al., 2025; Zhou et al., 2025; Huang et al., 2025; Li et al., 2025a).

Over-refusal in safety-aligned instruction-tuned models. Another line of work mitigates over-refusal in safety-aligned models through post-hoc model editing (Dabas et al., 2025) or inference-time

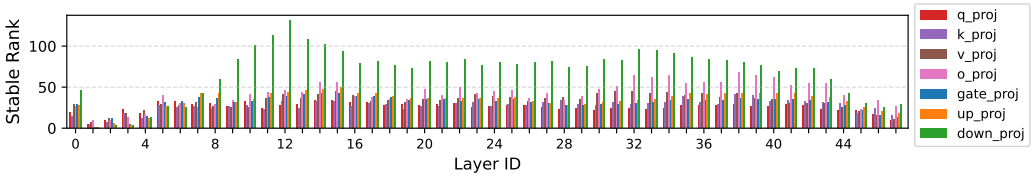


Figure 2: We compute the stable rank of the difference between the full-model finetuned model's weights and those of the original `DeepSeek-R1-Distill-Qwen-14B` for each layer. Here, the colors indicate the module types, and the x-axis shows the layer indices. We observe that the stable rank is quite high—ranging from around 40 to 100 for most layers

methods (Ray & Bhalani, 2024; Shi et al., 2024; Wang et al., 2024a). These methods address excessive refusal of benign requests rather than the degradation of reasoning caused by safety finetuning, and primarily target instruction-following behavior. We show that LoRA substantially lowers over-refusal.

Adversarial attacks and defenses. Adversarial red-teaming studies jailbreak prompts, intent-laundering obfuscation, and related attacks, often using benchmarks such as HarmBench (Mazeika et al., 2024). Defenses include prompt rewriting (Jain et al., 2023; Xie et al., 2023; Wang et al., 2024b), external safety classifiers (Alon & Kamfonas, 2023; Sharma et al., 2025), and decoding-time controls (Xu et al., 2024; Zou et al., 2023a). Our primary focus is refusal-based safety; but in Sec. 4, we show that low-rank adaptation also benefits adversarially robust safety finetuning.

Safety alignment of reasoning models. A trained model can be safety aligned through supervised finetuning (SFT) and/or reinforcement learning (RL) (Wei et al., 2021; Griffith et al., 2013; Dai et al., 2023; Ouyang et al., 2022; Rafailov et al., 2023; Bai et al., 2022; Guan et al., 2024). For reasoning models, Huang et al. (2025) identifies a key trade-off, termed the "Safety Tax," where safety alignment can substantially degrade reasoning ability. Jiang et al. (2025) construct a safety finetuning dataset with long CoT responses, but the resulting models still exhibit noticeable reasoning degradation (Huang et al., 2025). Existing work thus establishes the safety–reasoning trade-off but provides limited understanding of when safety alignment can preserve reasoning or why low-rank adaptation is effective in this setting.

In this work, we study why low-rank adaptation can mitigate this trade-off. We show that LoRA can effectively align reasoning models for refusal-based safety while largely preserving reasoning performance, and provide a theoretical explanation based on the interaction between the rank of the safety update and the intrinsic dimensionality of the capability being preserved. Our setting—***safety alignment of reasoning models***—is distinct from utility finetuning of an already safety-aligned model or safety alignment of non-reasoning models: LoRA's effectiveness depends critically on this setup, being strongest when the safety objective is effectively low-rank while the capability being preserved has high intrinsic dimensionality (Sec. 3.1; Fig. 4).

Theoretical positioning: LoRA and base-task preservation. Related work studies LoRA expressivity (Malladi et al., 2023), optimization (Jang et al., 2024), and preventing forgetting (Magistri et al., 2024) but does not characterize base-task preservation or how degradation depends on rank overshoot and intrinsic dimensionality. To our knowledge, this analysis is the first to establish this connection.

3 Safety Alignment Without Compromising Reasoning

In this section, we investigate whether the "Safety Tax" (Huang et al., 2025) can be mitigated—i.e., whether a model can be safety-aligned without sacrificing reasoning ability. We follow the setup of (Huang et al., 2025), performing supervised fine-tuning on safety datasets that pair harmful requests with refusal responses.

Motivation. Our key observation is that full-model finetuning, as in (Huang et al., 2025), induces high-rank weight changes. As shown in Figure 2, weight updates have high stable rank in most layers. In contrast, prior work identifies safety-related behavior with a small number of dominant directions in activation space (Panickssery et al., 2023; Arditi et al., 2024) and weight space (Wei et al., 2024; Jain et al., 2024), suggesting that low-rank weight changes may suffice to induce safe behavior without substantially modifying the broader weight space. We therefore hypothesize that

the Safety Tax may arise in part from additional high-rank changes that interfere with the learned reasoning solution. We adopt LoRA (Hu et al., 2022) because, unlike other PEFT methods (e.g., Prefix-tuning and Adapters), its update has a well-defined rank r, directly instantiating the object in our theory (Sec. 3.1).

To formalize this intuition, we present a theoretical analysis that isolates the key tension in finetuning: adapting to a new task while preserving a base task. We show that when the finetuning target is low-rank and the base target has large intrinsic dimension, full-model finetuning can fit the finetuning task perfectly while substantially degrading the base task. In contrast, restricting updates to low rank can simultaneously fit the finetuning task and better preserve the base task.

3.1 Theoretical Setup

Data and model. Fine-tuning dynamics are often well approximated by local, near-linear behavior around the pretrained initialization, with limited curvature (Fort et al., 2020). To capture this setting, we study a family of regression tasks where the goal is to learn a linear map from $\mathbb{R}^d$ to $\mathbb{R}^k$. Each task is specified by $\mathcal{T} = (\mathcal{V}, s, \sigma, \boldsymbol{\Theta})$, where $\mathcal{V} \subseteq \mathbb{R}^d$ is a task-specific signal subspace, $s > 0$ is the signal scale, $\sigma > 0$ is the noise scale, and $\boldsymbol{\Theta} \in \mathbb{R}^{d\times k}$ is the target map. Data are generated by $\boldsymbol{x} = \boldsymbol{z}+\boldsymbol{\xi}$ with $\boldsymbol{z} \sim \mathcal{N}(0, s^2\boldsymbol{P}_{\mathcal{V}})$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\boldsymbol{P}_{\mathcal{V}^\perp})$, where $\boldsymbol{P}_{\mathcal{V}}$ denotes the orthogonal projection matrix onto $\mathcal{V}$ and $\boldsymbol{P}_{\mathcal{V}^\perp}$ its orthogonal complement, and labels are $\boldsymbol{y} = \boldsymbol{z}^\top\boldsymbol{\Theta}$, where $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\Theta} = \boldsymbol{\Theta}$. We consider linear predictors $f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{W}$ trained under the squared loss $L(\boldsymbol{W}, \mathcal{T}) = \mathbb{E}\big[\|\boldsymbol{x}^\top\boldsymbol{W} - \boldsymbol{y}\|^2\big]$. For comparability across tasks, we fix the label scale such that $\mathbb{E}\|\boldsymbol{y}\|^2 = C$ for a fixed constant $C > 0$.

Base vs. finetuning tasks. We consider a base task $\mathcal{T}_{\text{base}} = (\mathcal{V}_{\text{base}}, s_{\text{base}}, \sigma_{\text{base}}, \boldsymbol{\Theta}_{\text{base}})$ and a finetuning task $\mathcal{T}_{\text{ft}} = (\mathcal{V}_{\text{ft}}, s_{\text{ft}}, \sigma_{\text{ft}}, \boldsymbol{\Theta}_{\text{ft}})$, and assume the following:

- (A1) $\mathcal{V}_{\text{base}} \perp \mathcal{V}_{\text{ft}}$, i.e., the signal subspaces of the base and finetuning tasks are orthogonal.
- (A2) $0 < \sigma_{\text{base}} = o(s_{\text{ft}})$ and $0 < \sigma_{\text{ft}} = o(s_{\text{base}})$, i.e., each task has nonzero noise, but its noise is asymptotically much smaller than the other task's signal.

Additionally, the following spectral separation condition is used only for Theorem 3.3:

- (A3) $\frac{\sigma_1(\boldsymbol{\Theta}_{\text{base}})}{\sigma_{r_{\text{ft}}}(\boldsymbol{\Theta}_{\text{ft}})} < \frac{s_{\text{ft}}}{\sigma_{\text{ft}}}$, where $\sigma_i(\cdot)$ denotes the i-th singular value in nonincreasing order. In words, the ratio between the largest singular value of $\boldsymbol{\Theta}_{\text{base}}$ and the smallest nonzero singular value of $\boldsymbol{\Theta}_{\text{ft}}$ is strictly smaller than the signal-to-noise ratio of the finetuning task. Intuitively, this is a signal-separation condition: it requires the weakest nonzero direction of the finetuning target to remain distinguishable, in the finetuning data, from the strongest component of the base task.

Finetuning from a base model. Finetuning initializes at $\boldsymbol{W}_0 = \boldsymbol{\Theta}_{\text{base}}$, which is optimal for the base task, and trains the model on the finetuning task. Next, we compare full-model and LoRA finetuning.

(i.) *Full-model fine-tuning.* We run gradient descent over $\boldsymbol{W}$ to minimize $L(\boldsymbol{W}, \mathcal{T}_{\text{ft}})$. Since the population loss is a convex quadratic function of $\boldsymbol{W}$ with a unique minimizer, for a sufficiently small stepsize $0 < \eta < \frac{1}{\max\{s_{\text{ft}}^2, \sigma_{\text{ft}}^2\}}$, gradient descent converges linearly to the following unique minimizer, which we treat as the final model obtained by full-model finetuning:

$$\boldsymbol{W}_{\text{full}} \in \arg\min_{\boldsymbol{W}} L(\boldsymbol{W}, \mathcal{T}_{\text{ft}}),$$

(ii.) *LoRA finetuning.* We reparameterize the model weights as $\boldsymbol{W} = \boldsymbol{W}_0 + \boldsymbol{B}\boldsymbol{A}$, where $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ are trainable parameters, and perform gradient descent over $(\boldsymbol{B}, \boldsymbol{A})$:

$$(\boldsymbol{B}_{t+1}, \boldsymbol{A}_{t+1}) = (\boldsymbol{B}_t, \boldsymbol{A}_t) - \eta\,\nabla_{\boldsymbol{B},\boldsymbol{A}} L(\boldsymbol{W}_0 + \boldsymbol{B}_t\boldsymbol{A}_t, \mathcal{T}_{\text{ft}}).$$

Under mild conditions on the stepsize and initialization (see detailed discussion in Appendix D.7), gradient descent converges to a global minimizer. We therefore focus on the converged parameters $(\boldsymbol{B}_\infty, \boldsymbol{A}_\infty)$. Equivalently, letting $\boldsymbol{S} := \boldsymbol{B}\boldsymbol{A}$, LoRA restricts the update to satisfy $\mathbf{rank}\boldsymbol{S} \le r$. Thus, the final weights obtained by LoRA can be written as

$$\boldsymbol{W}_{\text{lora}} = \boldsymbol{W}_0 + \boldsymbol{S}^*, \text{where } \boldsymbol{S}^* \in \arg\min_{\mathbf{rank}\boldsymbol{S}\le r} L(\boldsymbol{W}_0 + \boldsymbol{S}, \mathcal{T}_{\text{ft}}).$$

Key dimensions. We write $r_{\text{ft}} := \mathbf{rank}(\boldsymbol{\Theta}_{\text{ft}})$ and define the *intrinsic dimension* (stable rank):

$$\mathbf{intdim}(\boldsymbol{M}) \;:=\; \|\boldsymbol{M}\|_F^2/\|\boldsymbol{M}\|_2^2 \;\in [1, \mathbf{rank}(\boldsymbol{M})].$$

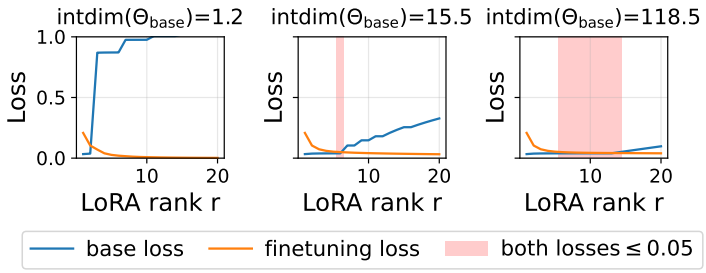


Figure 3: Numerical experiments under the theoretical setup. Singular values of $\mathbf{\Theta}_{\text{base}}$ and $\mathbf{\Theta}_{\text{ft}}$ follow a power-law decay. $\textbf{intdim}(\mathbf{\Theta}_{\text{ft}})$ is fixed at 1.2, while $\textbf{intdim}(\mathbf{\Theta}_{\text{base}})$ increases from left to right (1.2, 15.5, and 118.5). The red shaded region indicates values of r for which both losses are ≤ 0.05. As $\textbf{intdim}(\mathbf{\Theta}_{\text{base}})$ grows, this region expands from nonexistent to wide.

Overshoot. We also define the *overshoot* as $\Delta := (r - r_{\text{ft}})_+$, measuring how much the LoRA budget exceeds the minimal rank needed to represent the finetuning map exactly.

3.2 Comparison and a Deep Dive into LoRA

Full-model finetuning forgets the base task. First, we show that full-model finetuning achieves the finetuning optimum but overwrites the base solution.

Theorem 3.1 (Full-model finetuning forgets the base task)**.** *Assume (A1)–(A2). As proved in Appendix D.3, full finetuning satisfies:*

$$L(\boldsymbol{W}_{\text{full}}, \mathcal{T}_{\text{ft}}) = 0, \qquad L(\boldsymbol{W}_{\text{full}}, \mathcal{T}_{\text{base}}) \geq C.$$

LoRA yields satisfactory performance for low-rank finetuning tasks. Next, we show that the finetuning loss cannot be small unless the intrinsic dimension of $\mathbf{\Theta}_{\text{ft}}$ is comparable to the chosen rank r. In this case, LoRA can perform well. The proof is in Appendix D.4.

Theorem 3.2. *Assume (A1). For any LoRA solution with r,*

$$L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{ft}}) \;\geq\; C\Big(1 - \frac{r}{\textbf{\textit{intdim}}(\mathbf{\Theta}_{\text{ft}})}\Big)_+.$$

Consequently, if $\textbf{\textit{intdim}}(\mathbf{\Theta}_{\text{ft}}) \gg r$, then $L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{ft}}) \gtrsim C$.

Interaction between overshoot and the intrinsic dimension of the base task. When r is large enough to represent the finetuning map, LoRA can fit the finetuning task. However, the degree of base-task degradation depends on the overshoot Δ and on how spread out the base task is in parameter space. A large $\textbf{intdim}(\mathbf{\Theta}_{\text{base}})$ means the energy of $\mathbf{\Theta}_{\text{base}}$ is distributed across many directions, making the base task more tolerant to overshoot. The next result formalizes this: once r is large enough to fit the finetuning map, the finetuning loss becomes negligible, and the base loss scales inversely with $\textbf{intdim}(\mathbf{\Theta}_{\text{base}})$. The proof is in Appendix D.5.

Theorem 3.3 (Near-perfect LoRA finetuning when the base task has large intrinsic dimension)**.** *Assume (A1)–(A3). If $r \geq r_{\text{ft}}$, then the LoRA-finetuned model satisfies:*

$$L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{ft}}) = o(1), \qquad L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{base}}) \leq C \cdot \frac{\Delta}{\textbf{\textit{intdim}}(\mathbf{\Theta}_{\text{base}})} + o(1).$$

Consequently, if $\textbf{\textit{intdim}}(\mathbf{\Theta}_{\text{base}}) = \omega(\Delta)$, then $L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{base}}) = o(1)$ even with overshoot.

Remark 3.4 (Lower bounds and tightness)**.** The following rank-based lower bound always holds:

$$L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{base}}) \;\geq\; C \cdot \frac{\Delta}{\textbf{rank}(\mathbf{\Theta}_{\text{base}})} + o(1).$$

Moreover, under a mild flat-spectrum condition (see Appendix D.6), one obtains a matching lower bound of order

$$L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{base}}) \;\gtrsim\; C\,\frac{\Delta}{\textbf{intdim}(\mathbf{\Theta}_{\text{base}})}.$$

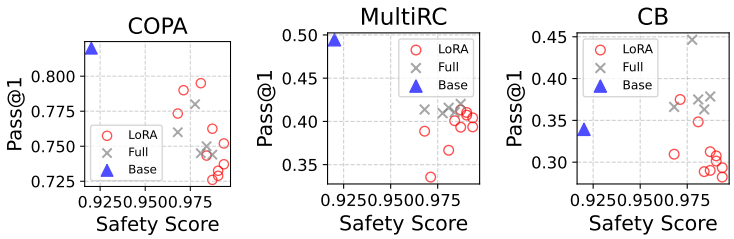


Figure 4: Safety finetuning on `Qwen2-1.5B-Instruct`. Utility is evaluated on instruction-following tasks (`COPA`, `MultiRC`, `CB`), and safety on `StrongREJECT`. In this non-reasoning, instruction-following setting, LoRA no longer provide an effective safety–utility trade-off.

3.3 Discussion: When LoRA is Effective and How to Tune It?

How to chose the LoRA rank? Theorem 3.3 and Remark 3.4 show that the sensitivity to overshoot Δ is governed by $\mathbf{intdim}(\mathbf{\Theta}_{\text{base}})$. In practice, it is rarely possible to choose the LoRA rank r perfectly, and since r is discrete, any mismatch yields $\Delta \geq 1$. Therefore, a key requirement for robust performance is that the base task have sufficiently large intrinsic dimension, so that it can tolerate inevitable overshoot. To illustrate this effect, we conduct numerical experiments that mirror our theoretical setup, setting the singular values of $\mathbf{\Theta}_{\text{base}}$ and $\mathbf{\Theta}_{\text{ft}}$ to follow a power-law decay. Fig. 3 plots the base and finetuning losses as a function of the LoRA rank r across three regimes, with the intrinsic dimension of the base task increasing from left to right. In the leftmost panel ($\mathbf{intdim}(\mathbf{\Theta}_{\text{base}}) = 1.2$), no choice of r achieves good performance on both tasks. In the middle panel ($\mathbf{intdim}(\mathbf{\Theta}_{\text{base}}) = 15.5$), good performance is achievable, but the base loss is highly sensitive and increases sharply even with a small overshoot; correspondingly, the red "good" region is narrow and requires a very precise choice of r. In the rightmost panel ($\mathbf{intdim}(\mathbf{\Theta}_{\text{base}}) = 118.5$), the good region widens substantially, demonstrating increased tolerance to overshoot and making LoRA much less sensitive to the exact choice of r.

Safety alignment with LoRA succeeds with high-rank base tasks and low-rank finetuning tasks. Theorems 3.2 and 3.3 formalize two complementary conditions under which LoRA is effective: the finetuning task must be sufficiently low-rank (otherwise the finetuning loss is large), and the base task must be sufficiently high-rank to tolerate the inevitable overshoot encountered in practice. In particular, safety finetuning a reasoning model matches this regime: the base capability we aim to preserve—*reasoning*—is highly complex and thus plausibly "high-rank" in nature. In contrast, the finetuning objective—*safety*—appears to be relatively low-rank, as suggested by prior work on steering vectors (Panickssery et al., 2023) and "refusal" features (Arditi et al., 2024; Yu et al., 2024), which show that safety-related behavior can be mediated by a single direction or a small number of directions in a model's internal representations. On the other hand, as instruction-tuning has low complexity, LoRA fails to provide a favorable safety–utility trade-off.

To probe this prediction in a more realistic setting, we consider safety finetuning an instruction-tuned (non-reasoning) model, where the base capability is instruction following which is intuitively less complex than reasoning. We finetune `Qwen2-1.5B-Instruct` for safety and evaluate utility on instruction-following tasks (Wang et al., 2019) including `COPA`, `MultiRC` and `CB`, and safety on `StrongREJECT`. Fig. 4 shows that LoRA is not effective in this regime and fails to provide a favorable safety–utility trade-off. This aligns with our theoretical insight that LoRA performs best when the base capability to be preserved is intrinsically high-dimensional, such as reasoning.

4 LoRA Bypasses the "Safety Tax"

This section introduces our safety finetuning and evaluation pipeline, then evaluates safety alignment and reasoning performance after full-model and LoRA safety finetuning.

4.1 Safety Alignment Fine-Tuning of Reasoning LLMs

We begin with a reasoning-capable language model and perform safety-alignment finetuning via SFT on a dataset of harmful questions paired with refusal responses, training the model to reject

harmful requests. In our training setup, we compare (1) **full-model finetuning**, as in (Huang et al., 2025), where all model parameters are updated using standard gradient-based optimization; and (2) **LoRA finetuning** described in Section 3.

4.2 Evaluation of the Finetuned Model

After safety alignment finetuning is completed, we evaluate two aspects of the model: (1) safety, assessed by sampling responses to a dataset of harmful questions and using `Llama-Guard-3-8B`—shown to be the strongest safety evaluator in (Jiang et al., 2025)—to judge whether each response is harmful; the *safety score* is the proportion of harmful responses. (2) reasoning ability, evaluated via Pass@1 on standard math, science, and coding benchmarks: for each question, we sample $n = 8$ responses and average the fraction judged correct.

4.3 Datasets and Models

Models. We consider three widely used open-weight reasoning models: `DeepSeek-R1-Distill-Qwen-7B`, `DeepSeek-R1-Distill-Qwen-14B`, and `DeepSeek-R1-Distill-Llama-8B`. These models span three sizes and two architectures (Qwen and Llama). Safety evaluation is performed using `Llama-Guard-3-8B`, which Jiang et al. (2025) found to be the most accurate safety evaluator.

Safety finetuning dataset. We use the `DirectRefusal` dataset, adapted from (Rosati et al., 2024) by (Huang et al., 2025), containing harmful requests with refusal answers.

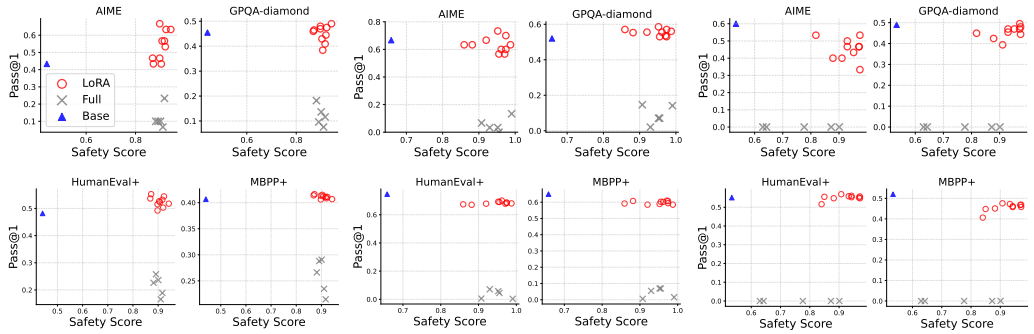
Safety evaluation datasets. We consider two datasets for safety evaluation: `StrongREJECT` (Souly et al., 2024), 310 policy-violating queries testing whether a model behaves safely, and `BeaverTails` (Ji et al., 2023), harmful data spanning 14 harm categories. All results in the main paper are reported on `StrongREJECT`, and we defer results on `BeaverTails` to Appendix B.1. The patterns observed in our experiments are consistent across both datasets. We focus not only on directly refusing harmful requests, but also on robustness to adversarial red-teaming attacks (e.g., jailbreak), evaluated via `GCG` and `AutoDAN` attacks on `StrongREJECT` prompts (Section 4, Appendix B.2), and on over-refusal, evaluated via `XSTest` (Röttger et al., 2023) (Appendix A.3). All safety judgments use `Llama-Guard-3-8B`, which we validate against human annotation on a sample of responses (Appendix A).

Reasoning benchmarks. We evaluate the models performance on (1) American Invitational Mathematics Examination 2024 (`AIME`), (2) `GPQA` (Rein et al., 2024) evaluate mathematical and scientific reasoning, respectively, (3) `HumanEval` (Chen et al., 2021) and (4) `MBPP` (Austin et al., 2021) are code generation benchmarks. We also consider the augmented versions created by `EvalPlus` (Liu et al., 2023), denoted as `HumanEval+` and `MBPP+`.

Training Setup. Full-model finetuning is performed for 5 epochs and LoRA is run for 10 epochs. We save and evaluate checkpoints at every epoch. Unless otherwise stated, LoRA is applied only to the MLP layers with rank $r = 4$. In Section 5, we investigate the effect of varying r and applying LoRA to different modules (e.g., MLP vs. attention) and layers. Additional details are deferred to Appendix A.

4.4 LoRA for Refusal-based Safety Alignment of Reasoning LLMs

Figure 5 compares safety and reasoning performance across checkpoints (epochs) during full-model and LoRA-based safety alignment. The base model achieves high reasoning accuracy but low safety, while full-model finetuning substantially improves safety at the cost of degraded reasoning. LoRA largely avoids this trade-off, maintaining strong performance on both safety and reasoning, as shown by the red points in the upper-right region. This pattern holds across three model sizes, two architectures, two safety benchmarks, and four reasoning benchmarks spanning mathematics, science, and code generation. In some settings, such as the 14B model on coding tasks and the 8B model on `AIME`, LoRA causes a slight drop in reasoning performance; in others, such as the 7B model, it even improves reasoning. Overall, the trend is consistent and demonstrates the effectiveness of LoRA for safety alignment.



(a) `R1-Distill-Qwen-7B` (b) `R1-Distill-Qwen-14B` (c) `R1-Distill-Llama-8B`

Figure 5: Across different model sizes and architectures, LoRA bypasses the "Safety Tax", achieving safety comparable to full-model finetuning while preserving reasoning performance close to that of the original reasoning model. We plot reasoning performance—measured by Pass@1—against safety scores for different models. For finetuned models, we report results for checkpoints from all epochs. Results on the base versions of `HumanEval` and `MBPP` are shown in Figure 10; the same pattern holds, but with higher accuracy. All safety scores here are measured on `StrongREJECT`; results on `BeaverTails` are deferred to Figure 8, where we observe the same pattern.

LoRA also reduces over-refusal. We further evaluate over-refusal on `XSTest` (Röttger et al., 2023) in Appendix A.3. At matched safety levels, LoRA has lower over-refusal than full-model finetuning across all three models and meets the safety floor at a substantially higher fraction of checkpoints.

LoRA also bypasses the Safety Tax under adversarial robustness. The results above consider refusal-based safety alignment without adversarial attacks such as jailbreaks. As a preliminary extension, we replace the safety objective with an adversarially robust one using continuous adversarial training (CAPO/CAT) (Xhonneux et al., 2024), while keeping the rest of the pipeline unchanged. We compare full-model finetuning with LoRA ($r = 4$) on `DeepSeek-R1-Distill-Qwen-7B`. Figure 6 shows the same pattern as Figure 5: at matched safety, full-model finetuning substantially degrades reasoning on AIME and GPQA, whereas LoRA better preserves reasoning. Under unseen jailbreak attacks (`GCG`, `AutoDAN`), LoRA also achieves robustness comparable to full-model finetuning (Appendix B.2, Table 2). These results suggest that the Safety Tax and the benefits of low-rank adaptation extend beyond refusal-based safety to adversarially robust objectives, while a more comprehensive study remains future work.

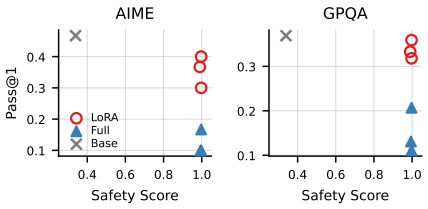


Figure 6: LoRA again bypasses the Safety Tax under an adversarially robust objective (CAPO/CAT): at matched safety, it preserves more reasoning than full-model finetuning, with comparable robustness to unseen jailbreaks (GCG, AutoDAN). Preliminary result on `DeepSeek-R1-Distill-Qwen-7B`; see Appendix B.2.

5 How Much LoRA is Enough?

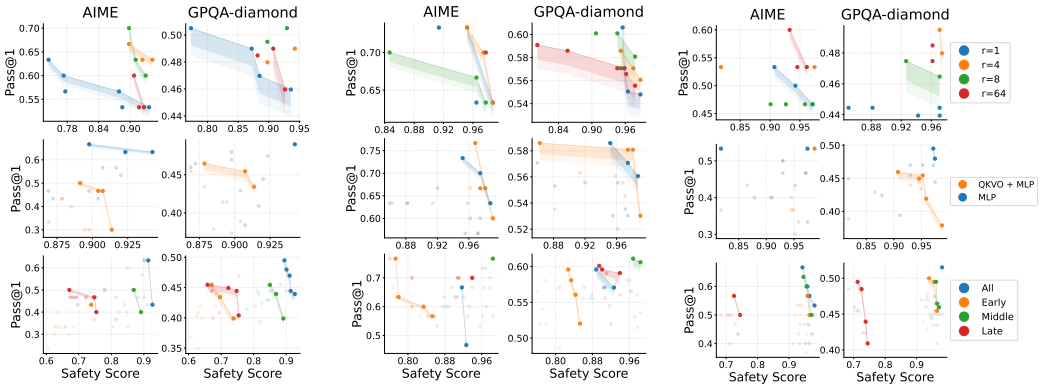
In this section, we ablate the LoRA configuration—(1) the rank r, (2) the modules, and (3) the layers to which LoRA is applied—to determine the smallest update sufficient to achieve both strong reasoning and safety. Unless otherwise noted, all ablations use the same three DeepSeek-distilled reasoning models as our main experiments: `R1-Distill-Qwen-7B`, `R1-Distill-Qwen-14B`, and `R1-Distill-Llama-8B`.

For each configuration, we train for 10 epochs and visualize the *Pareto frontier* of the resulting checkpoints in the reasoning–safety plane.

5.1 Rank: $r = 4$ is Sufficient

We explore the effect of the LoRA rank r by applying LoRA to MLP layers and varying r. From the first row of Figure 7, we observe that small ranks are sufficient, with $r = 4$ providing a strong default across models: for the larger 14B model, with higher parameter capacity, $r = 1$ achieves an excellent

balance, while $r = 4$ provides the best overall trade-off for the smaller models. These results suggest that the safety alignment task is effectively low-rank. Indeed, prior work has shown that safety behavior can often be mediated by a single direction in a model's internal representations, commonly referred to as a steering vector (Panickssery et al., 2023) or refusal features (Arditi et al., 2024; Yu et al., 2024), which may explain why low-rank updates suffice for effective safety alignment.



(a) `R1-Distill-Qwen-7B` (b) `R1-Distill-Qwen-14B` (c) `R1-Distill-Llama-8B`

Figure 7: Ablation results across three DeepSeek-distilled reasoning models (a)–(c). Axes show safety score (x) vs. Pass@1 (y); solid markers show the Pareto frontier over the 10-epoch training trajectory. **Row 1:** rank r—very small ranks suffice, with $r = 4$ a strong default and $r = 1$ already working well for the 14B model. **Row 2:** module choice at $r = 4$—MLP-only outperforms MLP+QKVO. **Row 3:** layer location—middle layers give the best trade-off.

5.2 Modules: MLP-Only LoRA Works Best

The LoRA adapter is usually applied to attention layers and/or MLP modules, e.g., by setting `target_modules` when using the `PEFT` package. Here, we compare two module configurations under the same rank setting ($r = 4$): applying LoRA only to MLP layers versus applying it to both attention and MLP layers (MLP+QKVO).

From the second row of Figure 7, we observe that MLP-only LoRA consistently gives a better reasoning–safety trade-off than MLP+QKVO. In other words, adding QKVO adapters is not necessary in our setup and can even hurt the overall Pareto frontier.

5.3 Layers: Middle Layers Matter Most

We ablate layer locations across the three models, comparing all, early, middle, and late layers; the precise ranges differ by model and are given in Appendix A. In all cases, we apply LoRA only to the up-projection layers with $r = 4$.

From the third row of Figure 7, applying LoRA to middle layers generally yields the best reasoning–safety trade-off, while early or late layers produce a noticeably worse frontier. Middle-layer LoRA also performs close to using all layers in many cases, suggesting that safety alignment is concentrated in a relatively localized portion of the network.

This is consistent with prior findings on steering vectors (Panickssery et al., 2023) and refusal features (Arditi et al., 2024; Yu et al., 2024), which suggest that safety-related directions in intermediate representations are most prominent in middle layers.

6 Conclusion

Safety alignment of reasoning LLMs presents a fundamental challenge: improving safety without sacrificing the reasoning capabilities acquired during training. We show that LoRA effectively addresses this challenge, achieving safety comparable to full-model alignment while largely preserving reasoning and reducing over-refusal across diverse models and benchmarks. Our theoretical analysis explains when and why low-rank adaptation is effective, linking base-task degradation to rank overshoot and the intrinsic dimensionality of the capability being preserved. These results establish low-rank adaptation as a simple and effective approach for safety alignment of reasoning models.

AI USE STATEMENT

We used generative AI tools to polish the writing (grammar, phrasing, and clarity) in this manuscript, and to assist in writing portions of the experiment and data-processing scripts used in our codebase. We did not use generative AI tools to generate research ideas, theoretical claims, or experimental results. All AI-assisted text was reviewed and edited by the authors, and all AI-assisted code was reviewed and tested by the authors for correctness before use. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies safety alignment of reasoning models and, as such, uses existing public datasets containing harmful requests (e.g., `DirectRefusal`, `StrongREJECT`, `BeaverTails`) that are standard in the safety alignment literature. We do not collect new harmful content, involve human subjects, or release any new sensitive data; our contribution is a training recipe intended to improve model safety, and we do not foresee additional ethical concerns beyond those already associated with the public datasets and models we build on.

REPRODUCIBILITY STATEMENT

We plan to release our training and evaluation code publicly upon publication. Complete statements of our assumptions and full proofs of Theorems 3.1–3.3 are provided in Appendix D. Full experimental details, including dataset processing, training hyperparameters, and evaluation protocols, are provided in Appendix A.

REFERENCES

Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.

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. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Hyeong Kyu Choi, Xuefeng Du, and Yixuan Li. Safety-aware fine-tuning of large language models. *arXiv preprint arXiv:2410.10014*, 2024.

Mahavir Dabas, Si Chen, Charles Fleming, Ming Jin, and Ruoxi Jia. Just enough shifts: Mitigating over-refusal in aligned language models with targeted representation fine-tuning. *arXiv preprint arXiv:2507.04250*, 2025.

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.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL `https://zenodo.org/records/12608602`.

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26, 2013.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Lei Hsiung, Tianyu Pang, Yung-Chen Tang, Linyue Song, Tsung-Yi Ho, Pin-Yu Chen, and Yaoqing Yang. Why llm safety guardrails collapse after fine-tuning: A similarity analysis between alignment and fine-tuning datasets. *arXiv preprint arXiv:2506.05346*, 2025.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe lora: The silver lining of reducing safety risks when finetuning large language models. *Advances in Neural Information Processing Systems*, 37:65072–65094, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

Yeonjun In, Wonjoong Kim, Sangwu Park, and Chanyoung Park. Reasoning structure matters for safety alignment of reasoning models. *arXiv preprint arXiv:2604.18946*, 2026.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Samyak Jain, Ekdeep S Lubana, Kemal Oksuz, Tom Joy, Philip Torr, Amartya Sanyal, and Puneet Dokania. What makes and breaks safety fine-tuning? a mechanistic study. *Advances in Neural Information Processing Systems*, 37:93406–93478, 2024.

Uijeong Jang, Jason D Lee, and Ernest K Ryu. Lora training in the ntk regime has no spurious local minima. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 21306–21328, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*, 2025.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

Ang Li, Yichuan Mo, Mingjie Li, Yifei Wang, and Yisen Wang. Are smarter llms safer? exploring safety-reasoning trade-offs in prompting and fine-tuning. *arXiv preprint arXiv:2502.09673*, 2025a.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*, 2025b.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572, 2023.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *arXiv preprint arXiv:2402.18540*, 2024.

Simone Magistri, Joost van de Weijer, Andew D Bagdanov, et al. An empirical analysis of forgetting in pre-trained models with incremental low-rank updates. *arXiv preprint arXiv:2405.18069*, 2024.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pp. 23610–23641. PMLR, 2023.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Jishnu Mukhoti, Yarin Gal, Philip HS Torr, and Puneet K Dokania. Fine-tuning can cripple your foundation model; preserving features may be the solution. *arXiv preprint arXiv:2308.13320*, 2023.

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.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

ShengYun Peng, Pin-Yu Chen, Jianfeng Chi, Seongmin Lee, and Duen Horng Chau. Shape it up! restoring llm safety during finetuning. *arXiv preprint arXiv:2505.17196*, 2025.

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! *arXiv preprint arXiv:2310.03693*, 2023.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Ruchira Ray and Ruchi Bhalani. Mitigating exaggerated safety in large language models. *arXiv preprint arXiv:2405.05418*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, Subhabrata Majumdar, Hassan Sajjad, Frank Rudzicz, et al. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676, 2024.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. *arXiv preprint arXiv:2410.07471*, 2024.

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuan-Jing Huang, Xun Zhao, and Dahua Lin. Navigating the overkill in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4602–4614, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Hossein Valavi, Sulin Liu, and Peter Ramadge. Revisiting the landscape of matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1629–1638. PMLR, 2020.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. *arXiv preprint arXiv:2410.03415*, 2024a.

Yihan Wang, Zhouxing Shi, Andrew Bai, and Cho-Jui Hsieh. Defending llms against jailbreaking attacks via backtranslation. *arXiv preprint arXiv:2402.16459*, 2024b.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, and Gauthier Gidel. Efficient adversarial training in LLMs with continuous attacks. *arXiv preprint arXiv:2405.15589*, 2024.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5 (12):1486–1496, 2023.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. *arXiv preprint arXiv:2402.08983*, 2024.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust llm safeguarding via refusal feature adversarial training. *arXiv preprint arXiv:2409.20089*, 2024.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*, 2023.

Weixiang Zhao, Xingyu Sui, Jiahe Guo, Yulin Hu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, Tat-Seng Chua, and Ting Liu. Trade-offs in large reasoning models: An empirical analysis of deliberative and adaptive reasoning over foundational capabilities. *arXiv preprint arXiv:2503.17979*, 2025.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. The hidden risks of large reasoning models: A safety assessment of r1. *arXiv preprint arXiv:2502.12659*, 2025.

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2):1308–1331, 2021.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

A Experimental Details

A.1 Details about DirectRefusal

The original `DirectRefusal` dataset can be downloaded from `TianshengHuang/DirectRefusal` on Hugging Face. For each harmful request, the corresponding response includes a short thought—"I should not answer this question!"—followed by a refusal answer. We adapted the dataset to ensure it follows the `DeepSeek-R1` chat template.

A.2 Details about Safety Evaluation

For each question, we sample the answer using greedy decoding and pass the answer to `Llama-Guard-3-8B` to judge whether it is safe. Safety is determined by checking whether the substring "safe" or "unsafe" is contained in the evaluator's output. We use `Llama-Guard-3-8B` because Jiang et al. (2025) found it to be the strongest and most accurate safety evaluator for this type of model among those they benchmarked.

To check its reliability, we manually re-annotated a sample of 30 responses and compared the labels against `Llama-Guard`'s judgments. Of the 28 responses a human labeled safe, `Llama-Guard` agreed on 26 and flagged 2 as unsafe; of the 2 responses a human labeled unsafe, `Llama-Guard` agreed on both. Overall agreement is 28/30 (93.3%), and `Llama-Guard` did not miss any response a human annotator judged unsafe (zero false negatives)—its only disagreements are in the conservative direction, flagging as unsafe two responses a human judged safe.

A.3 Over-Refusal Evaluation on XSTest

To evaluate over-refusal, we use `XSTest` (Röttger et al., 2023), which pairs genuinely harmful prompts with superficially similar but benign prompts that a well-calibrated model should answer. For each checkpoint, we report two rates: *over-refuse*, the refusal rate on benign prompts that should be answered, and *should-refuse*, the refusal rate on genuinely harmful prompts. Refusal is judged using the same `Llama-Guard-3-8B` procedure as our main safety evaluation.

For each model and finetuning method, we select the checkpoint with the lowest over-refuse rate among those meeting a safety floor of should-refuse ≥ 0.85, and report the fraction of checkpoints (out of all epochs trained) that meet this floor. Table 1 summarizes the results.

Table 1: Over-refusal on `XSTest`. At a matched safety level (should-refuse ≥ 0.85), LoRA achieves a lower over-refuse rate than full-model finetuning on all three models, and meets the safety floor at a higher fraction of its checkpoints.

Model	Method	Epoch	Over-refuse ↓	Should-refuse ↑	Checkpoints meeting floor
Qwen-7B	Full	5	0.620	0.860	4/5
Qwen-7B	LoRA	3	0.568	0.860	8/10
Qwen-14B	Full	5	0.664	0.865	3/5
Qwen-14B	LoRA	3	0.248	0.860	5/10
Llama-8B	Full	5	0.760	0.895	3/5
Llama-8B	LoRA	3	0.328	0.870	10/10

LoRA meeting the safety floor at a substantially higher fraction of its checkpoints (e.g., 10/10 for Llama-8B versus 3/5 for full finetuning) indicates that its favorable safety–reasoning trade-off is stable across training, not just achievable at a single cherry-picked checkpoint.

The over-refusal rates we observe for full-model finetuning are consistent with those reported elsewhere for the same finetuning dataset and benchmark: In et al. (2026) evaluate full finetuning on `DirectRefusal` using `XSTest`, reporting over-refusal rates of 14.8–72.4 across model scales 1.5B–32B. `XSTest` attributes such over-refusal to "lexical overfitting"—models trained on refusal data become oversensitive to safety-related words that predominantly co-occur with harmful content in training (Röttger et al., 2023)—and Qi et al. (2024) show that safety alignment trained on short, templated refusal responses, exactly `DirectRefusal`'s structure, produces "shallow" alignment that pattern-matches on refusal prefixes rather than reasoning about context. In et al. (2026) further show this is fixable by scaling the training set (from 1K to 3K examples, reducing over-refusal from 31.6 to 2.4 on R1-7B), indicating that over-refusal severity tracks training data size and diversity rather than being an inherent property of any one finetuning mechanism. Our results show that, independent of this data-side effect, LoRA itself already reduces over-refusal relative to full-model finetuning on the same dataset.

A.4 Details about `AIME` and `GPQA`

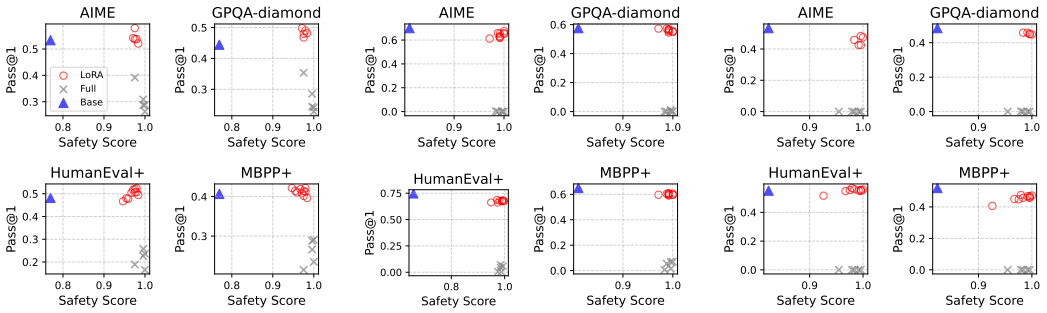
We adapted the codebase from Li et al. (2025b), which builds on top of `lm-evaluation-harness` Gao et al. (2024), to evaluate models on `GPQA` and `AIME`. For `AIME`, answers are evaluated using `Qwen2.5-32B-Instruct` as a judge. For `GPQA`, since it is a multiple-choice benchmark, answers are evaluated via regular expression matching. To improve evaluation accuracy, we made some small adjustments. One change is modifying the prompt to better ensure the model outputs answers in the desired format:

```
You are solving a multiple-choice question. At the end, present your
    final answer using the format: Final Answer: \boxed{X}, where X is
    one of A, B, C, or D.

Question: {{Question}}
Choices:
(A) {{choice1}}
(B) {{choice2}}
(C) {{choice3}}
(D) {{choice4}}
```

Another adjustment is in the answer extraction logic—we enhanced the original implementation to handle a wider range of answer formats that models produce.

During generation, we set the temperature to 0.6, top_p to 0.95, and the maximum number of generated tokens to 32,768.



(a) `R1-Distill-Qwen-7B` (b) `R1-Distill-Qwen-14B` (c) `R1-Distill-Llama-8B`

Figure 8: Results with safety measured on `BeaverTails`, consistent with the pattern shown in Figure 5.

A.5 Details about HumanEval and MBPP

We adapted the codebase of `EvalPlus` Liu et al. (2023). The original implementation included a response prefix, designed for earlier models that did not explicitly support an intermediate reasoning process. This prefix—such as "Below is a Python script with a self-contained function that efficiently solves the problem and passes corresponding tests:"—was prepended to model outputs during generation. However, this practice introduces unfair bias by encouraging all models to directly generate code. It inadvertently benefits the full-model fine-tuned baseline—which would otherwise often refuse to answer—by effectively forcing it to produce code. Conversely, it disadvantages reasoning-aligned models by disrupting the expected format that includes intermediate "thought", causing them to skip the thinking process entirely. This can result in skewed conclusions. We suspect this explains the abnormally low accuracy of the base reasoning model and the high accuracy of the full-model fine-tuned variant reported in Jiang et al. (2025). To address this issue, we remove the response prefix. We also slightly reword the prompt to better align with reasoning models.

During generation, we set the temperature to 0.6 and the maximum number of generated tokens to 32,768.

A.6 Training Details

For 7B models (Qwen) and 8B models (Llama), full-model finetuning is conducted using 4 GPUs with a batch size of 2 per device for 5 epochs. LoRA finetuning uses 2 GPUs with a batch size of 2 per device for 10 epochs. We set the LoRA hyperparameters as $\alpha = 16$ and `lora_dropout` = 0.05.

For 14B models, full-model finetuning is performed using 8 GPUs with a batch size of 1 per device for 5 epochs. LoRA finetuning uses 4 GPUs with a batch size of 2 per device for 10 epochs. We set the LoRA hyperparameters as $\alpha = 16$ and `lora_dropout` = 0.05.

Across all experiments, we use a learning rate of 5e-5 and a weight decay of 1e-4.

Layer ranges for the layer-location ablation (Section 5). For `R1-Distill-Qwen-7B`, we compare all layers (1–28), early layers (3–12), middle layers (10–19), and late layers (15–23). For `R1-Distill-Qwen-14B`, we compare all layers (1–48), early layers (5–20), middle layers (17–32), and late layers (25–40). For `R1-Distill-Llama-8B`, we compare all layers (1–32), early layers (3–13), middle layers (11–21), and late layers (17–27).

B Additional Plots for Experiments in Section 4

B.1 Results on BeaverTails

The results on `BeaverTails` across models are shown in Figure 8, and the observed pattern is consistent with that in Figure 5.

B.2 LoRA vs. Full-Model Finetuning under an Adversarially Robust Safety Objective

This section provides full details for the preliminary extension summarized in Section 4 (Figure 6), which asks whether LoRA continues to bypass the Safety Tax when the safety-finetuning objective is adversarially robust rather than plain refusal-based SFT.

Model and data. We use `DeepSeek-R1-Distill-Qwen-7B` and the same `DirectRefusal` harmful prompts used throughout the paper. All safety and robustness judgments use `Llama-Guard-3-8B`, the same evaluator used elsewhere in the paper.

Training objective. We replace plain refusal SFT with continuous adversarial training in the style of CAT/CAPO (Xhonneux et al., 2024). For each harmful prompt x from `DirectRefusal`, we construct a triple $(x, y, \hat{y})$, where y is the paired refusal response and $\hat{y}$ is a templated *affirmative* target (e.g., "Sure, here is ⟨restated request⟩:"; no real harmful content is generated or used). An inner attack finds a continuous perturbation δ on the prompt's token embeddings via signed-gradient ascent (10 steps, per-token ℓ_2 projection to a ball of radius $\varepsilon = 0.05\times$ the mean embedding norm) that maximizes $\log p_\theta(\hat{y} \mid x + \delta)$, i.e., simulating a worst-case embedding-space jailbreak. The outer loss then trains the model to still produce the refusal y under this perturbation,

$$\mathcal{L} = \mathrm{NLL}(y \mid x + \delta) + \lambda \cdot \mathrm{relu}\big(\overline{\log p_\theta}(\hat{y} \mid x + \delta) - c\big),$$

with $\lambda = 1.0$ and margin $c = -5.0$, discouraging the model from assigning high likelihood to the affirmative target even under attack. Critically, this objective requires no additional benign/utility data beyond `DirectRefusal` itself: unlike SFT-only baselines, the away term above already regularizes the model away from harmful completions without a separate utility loss.

Arms compared. *LoRA*: $r = 4$, MLP-only (`gate_proj`, `up_proj`, `down_proj`), $\alpha = 16$, dropout 0.05, trained for 10 epochs. *Full-model finetuning*: all parameters, trained for 3 epochs (safety converges quickly: training loss $1.06 \to 0.27 \to 0.08$ over epochs 1–3). Both arms use identical data, attack hyperparameters, and optimizer (AdamW, lr 5×10^{-5}, cosine schedule, bf16); the only difference is the set of parameters updated.

Evaluation. *Reasoning*: AIME 2024 and GPQA-Diamond, Pass@1. *Clean safety*: `StrongREJECT`, judged unsafe/safe by `Llama-Guard-3-8B`. *Adversarial robustness*: $1-$ attack success rate (ASR) under discrete jailbreak attacks *not seen during training*—`GCG` (Zou et al., 2023b) (200 optimization steps) and `AutoDAN` (Liu et al., 2024)—applied to a subset of `StrongREJECT` prompts, with success judged by the same `Llama-Guard-3-8B` evaluator. Training against a continuous attack and evaluating against unseen discrete attacks tests whether robustness generalizes, rather than merely fitting the training-time attack.

Results. Table 2 reports reasoning, clean safety, and robustness at each checkpoint. The same qualitative pattern as Figure 5 emerges: full-model finetuning and LoRA reach comparable robustness (GCG robustness in 0.93–1.00 for both), but full-model finetuning does so at a much larger cost to reasoning (AIME drops to 0.10–0.17, vs. 0.30–0.40 for LoRA). At matched clean safety (≈ 0.99), LoRA (epoch 6) preserves substantially more reasoning than any full-model finetuning checkpoint while remaining highly robust to both attacks. Figure 9 additionally breaks down robustness by attack type for representative checkpoints.

Scope and limitations. This is a single-model, preliminary study intended to test whether the Safety Tax phenomenon and LoRA's ability to avoid it extend to adversarially robust objectives, not a comprehensive robustness evaluation. We defer a broader study—additional models, ranks, and attacks (e.g., `PAIR`, reasoning-specific attacks such as `H-CoT`)—to future work.

B.3 Results on the Base Versions of HumanEval and MBPP

Figure 10 shows the results on the base versions of `HumanEval` and `MBPP`.

Table 2: LoRA vs. full-model finetuning under the adversarially robust (CAPO/CAT) safety objective, on `DeepSeek-R1-Distill-Qwen-7B`. Reasoning is Pass@1 (AIME: $n = 1$ greedy; GPQA-Diamond). Clean safety is 1−unsafe rate on `StrongREJECT`. GCG/AutoDAN robustness is 1−ASR under attacks unseen during training.

Method	AIME	GPQA-D	Clean Safety	GCG Rob.	AutoDAN Rob.
Base (no safety FT)	0.467	0.369	0.339	0.400	0.167
LoRA $r = 4$ (ep2)	0.400	0.359	0.997	0.967	1.000
LoRA $r = 4$ (ep4)	0.300	0.318	0.997	0.900	0.967
LoRA $r = 4$ (ep6)	0.367	0.333	0.990	0.967	0.967
Full FT (ep1)	0.167	0.207	0.997	1.000	1.000
Full FT (ep2)	0.100	0.131	0.994	0.967	0.967
Full FT (ep3)	0.100	0.111	0.997	0.933	1.000

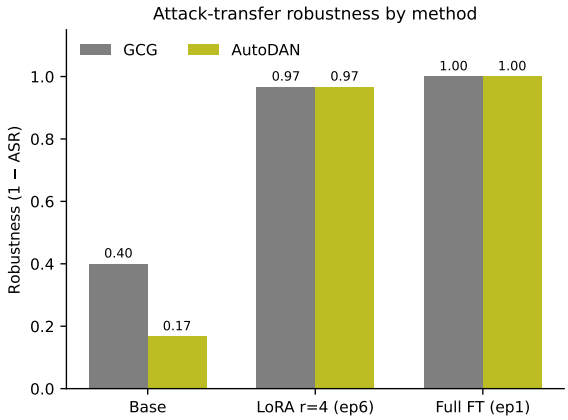
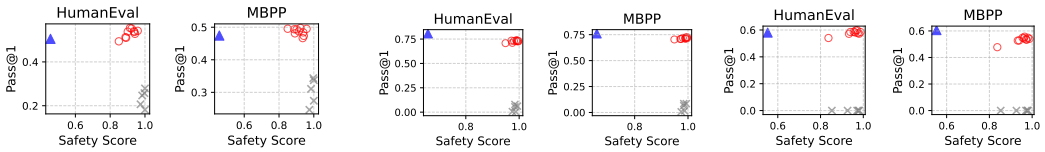


Figure 9: Robustness (1−ASR) under `GCG` and `AutoDAN` for the base model and representative LoRA/full-finetuning checkpoints (LoRA epoch 6, matched to full-finetuning epoch 1 by clean safety). Both finetuning methods substantially improve robustness over the base model against both attacks.

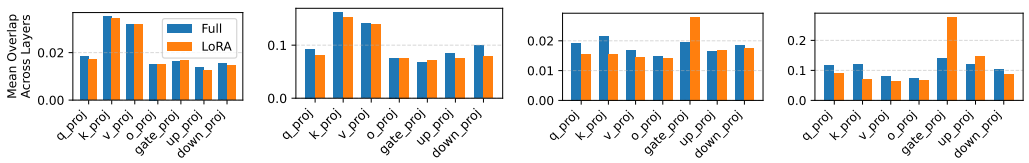


(a) `R1-Distill-Qwen-7B` (b) `R1-Distill-Qwen-14B` (c) `R1-Distill-Llama-8B`

Figure 10: Results on the base versions of `HumanEval` and `MBPP` show the same pattern as in the plus versions shown in Figure 5, but with higher accuracy.

C Additional Experimental Results: Exploring the Structure of LoRA Weights

We empirically study the structure of LoRA weight updates. Preliminary results suggest that LoRA induces updates with slightly smaller overlap with the initial weights than full-model finetuning, which may indicate that LoRA is less disruptive to reasoning-related weights. We then explore methods to explicitly reduce this overlap (via regularization or post-hoc weight merging). One post-hoc method yields a modest improvement in the reasoning–safety trade-off on some tasks. These results are preliminary, and further work is needed to develop approaches that produce consistent gains across tasks; we hope our findings will inspire future work in this direction.



(a) $\frac{\|\boldsymbol{W}_I^\top \Delta \boldsymbol{W}\|}{\|\boldsymbol{W}_I\|\|\Delta \boldsymbol{W}\|}$ (b) $\frac{\|\boldsymbol{U}_{16}\boldsymbol{U}_{16}^\top \Delta \boldsymbol{W}\|}{\|\Delta \boldsymbol{W}\|}$ (c) $\frac{\|\boldsymbol{W}_I \Delta \boldsymbol{W}^\top\|}{\|\boldsymbol{W}_I\|\|\Delta \boldsymbol{W}\|}$ (d) $\frac{\|\boldsymbol{V}_{16}\boldsymbol{V}_{16}^\top \Delta \boldsymbol{W}^\top\|}{\|\Delta \boldsymbol{W}\|}$

Figure 11: LoRA updates exhibit smaller overlap with the original weights compared to the full-model finetuning updates. Although the reduction in overlap is sometimes small, it can be observed across most layers for all four metrics, which cover both the column (a)(b) and row spaces (c)(d). The 14B models are used here.

C.1 LoRA Updates Have Less Alignment with Initial Weights

Intuitively, if the initial (reasoning) weights $\boldsymbol{W}_I$ and the LoRA update $\Delta \boldsymbol{W}$ have only a small alignment, this suggests minimal interference between the safety-oriented update and the weights critical for reasoning. LoRA already constrains $\Delta \boldsymbol{W}$ to be low-rank, meaning it spans only a small subspace of the full weight space. We further examine the orientation of this subspace: how do the directions learned by LoRA compare to those spanned by $\boldsymbol{W}_I$? To quantify this, we compute the following four metrics: (1) $\frac{\|\boldsymbol{W}_I^\top \Delta \boldsymbol{W}\|}{\|\boldsymbol{W}_I\|\|\Delta \boldsymbol{W}\|}$, (2) $\frac{\|\boldsymbol{U}_{16}\boldsymbol{U}_{16}^\top \Delta \boldsymbol{W}\|}{\|\Delta \boldsymbol{W}\|}$, (3) $\frac{\|\boldsymbol{W}_I \Delta \boldsymbol{W}^\top\|}{\|\boldsymbol{W}_I\|\|\Delta \boldsymbol{W}\|}$, and (4) $\frac{\|\boldsymbol{V}_{16}\boldsymbol{V}_{16}^\top \Delta \boldsymbol{W}^\top\|}{\|\Delta \boldsymbol{W}\|}$. Here, $\boldsymbol{U}_{16}$ and $\boldsymbol{V}_{16}$ are matrices containing the top 16 left and right singular vectors of $\boldsymbol{W}_I$, respectively, obtained via truncated SVD. Intuitively, (1) and (2) capture the overlap between $\boldsymbol{W}_I$ and $\Delta \boldsymbol{W}$ in the column space: (1) is a matrix-level analogue of cosine similarity, while (2) measures the normalized projection of $\Delta \boldsymbol{W}$ onto the top dominant directions of $\boldsymbol{W}_I$. Similarly, (3) and (4) capture overlap in the row space. The column and row spaces correspond to the directions the matrices "write to" and "read from", respectively. A smaller value in any of these metrics indicates greater orthogonality between $\boldsymbol{W}_I$ and $\Delta \boldsymbol{W}$ in the corresponding space.

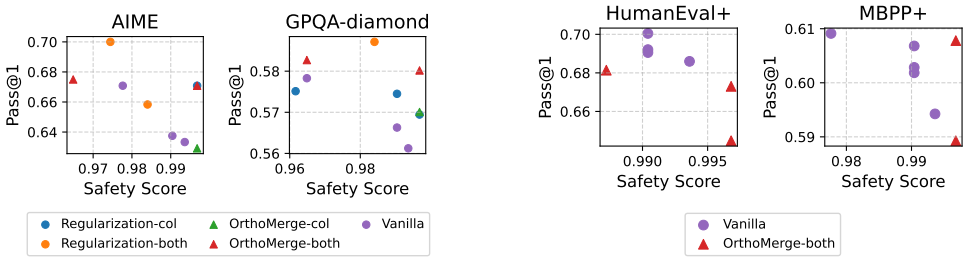
We compare the full-rank fine-tuned model with the LoRA fine-tuned model in which both the attention and MLP modules are updated with $r = 4$, making the two settings more comparable since updates occur in all major modules. We compute the alignment metrics for different module types across layers, average them over layers, and report the results for the 14B models in Figure 11. We observe that LoRA achieves smaller overlap in most modules across the metrics, with a few exceptions. This suggests that, for most weights, $\boldsymbol{W}_I$ and $\Delta \boldsymbol{W}$ are more orthogonal—both in the column and row spaces—for LoRA than for full-model finetuning. In other words, under LoRA finetuning, the safety-oriented updates read from and write into subspaces that are more separate from those used by the original reasoning-related weights—more so than in the full-model fine-tuned version. Although the reduction in alignment values is sometimes small, it may still indicate that LoRA updates interfere less with the reasoning-related components of the model, potentially explaining the better preservation of reasoning performance. A more in-depth investigation is needed to fully understand the underlying mechanisms and to develop more precise metrics for capturing this effect—an important direction for future work.

C.2 Exploring Methods That Further Reduce Alignment

Given the observations in Section C.1, we ask whether further reducing the overlap between $\Delta \boldsymbol{W}$ and $\boldsymbol{W}_I$ could lead to even better reasoning performance without compromising safety. This question is particularly relevant because, while the results achieved by LoRA in Section 4.4 are strong, they are not perfect—a small performance gap remains compared to the original reasoning model, especially for the 14B model on `AIME`, `HumanEval+`, and `MBPP+`.

We experiment with two approaches:

- **Regularization during LoRA training.** We add a penalty term to the loss that discourages overlap between $\boldsymbol{W}_I$ and $\Delta \boldsymbol{W}$ (to avoid out-of-memory issues during training caused by the



(a) 14B models on `AIME` and `GPQA`.

(b) 14B models on `HumanEval+` and `MBPP+`.

Figure 12: For each method that produces multiple checkpoints (e.g., across epochs or hyperparameter settings), we visualize the Pareto-frontier points. Methods enforcing orthogonality in both row and column spaces perform better than column-only variants. The post-hoc method `OrthoMerge-both` is the most promising, with points concentrated in the upper-right corner and its best point strictly dominating vanilla LoRA on `AIME` and `GPQA`. On `MBPP+`, it achieves a slightly better Pareto frontier, while on `HumanEval+` it slightly underperforms.

large dimensionality of model weights, in implementation we use a low-rank approximation of $\boldsymbol{W}_I$ instead of the full matrix), specifically:

  - `reg-col`: $\beta(\frac{\|\boldsymbol{W}_I^\top \Delta \boldsymbol{W}\|}{\|\boldsymbol{W}_I\|\|\Delta \boldsymbol{W}\|})^2$ encourages orthogonality in the column space.
  - `reg-both`: $\beta(\frac{\|\boldsymbol{W}_I^\top \Delta \boldsymbol{W}\|}{\|\boldsymbol{W}_I\|\|\Delta \boldsymbol{W}\|})^2 + \beta(\frac{\|\Delta \boldsymbol{W}^\top \boldsymbol{W}_I\|}{\|\boldsymbol{W}_I\|\|\Delta \boldsymbol{W}\|})^2$ encourages orthogonality in both column and row spaces.

  We tried different values of β but found no significant change in the results, so we fix $\beta = 1$.

- **Post-hoc Orthogonalization:** We post-process $\Delta \boldsymbol{W}$ before applying Equation **??**, projecting it onto the orthogonal complement of the top-k singular vectors of $\boldsymbol{W}_I$:

  - `OrthoMerge-col`: $\Delta \boldsymbol{W} \leftarrow (\boldsymbol{I} - \boldsymbol{U}_k \boldsymbol{U}_k^\top)\Delta \boldsymbol{W}$ enforces column-space orthogonality based on the rank-k SVD of $\boldsymbol{W}_I$.
  - `OrthoMerge-both`: $\Delta \boldsymbol{W} \leftarrow (\boldsymbol{I} - \boldsymbol{U}_k \boldsymbol{U}_k^\top)\Delta \boldsymbol{W}(\boldsymbol{I} - \boldsymbol{V}_k \boldsymbol{V}_k^\top)$ enforces orthogonality in both column and row spaces. We found that directly applying OrthoMerge-both leads to a drop in safety scores. To mitigate this, we further scale up the orthogonal complement with $\Delta \boldsymbol{W} \leftarrow \lambda(\boldsymbol{I} - \boldsymbol{U}_k \boldsymbol{U}_k^\top)\Delta \boldsymbol{W}(\boldsymbol{I} - \boldsymbol{V}_k \boldsymbol{V}_k^\top)$ to compensates for the loss in safety. We experiment with different values of λ in the range $\{1, 1.15, 1.75, 1.2, 1.25\}$. We set $k = 64$.

For both approaches, we omit the row-space-only variant, as it showed no significant improvement in our experiments.

Figure 12a shows the results on `AIME` and `GPQA` for the 14B models with $r = 4$. For each method that yields multiple checkpoints (e.g., different epochs for vanilla and regularization finetuning, or different hyperparameters for `OrthoMerge`), we visualize the Pareto-frontier points. We observe that methods addressing both the row and column spaces ("both") tend to yield better results than those that only operate on the column space ("col"). Among them, `OrthoMerge-both` appears most promising, with more points concentrated in the upper-right corner and its best point strictly dominating vanilla LoRA. Therefore, we additionally evaluate `OrthoMerge-both` on the coding benchmarks in Figure 12b, where it slightly underperforms on `HumanEval+` but slightly outperforms vanilla on `MBPP+` in terms of the tradeoff.

Overall, we observe modest yet inconsistent gains from post-hoc orthogonalization. This highlights the potential of controlling the subspace geometry of LoRA updates and points to the need for more nuanced methods that yield consistent improvements.

D Complete Proofs for Section 3.1

D.1 Basic Lemmas

Lemma D.1 (Population loss in closed form). *Fix a task $\mathcal{T} = (\mathcal{V}, s, \sigma, \boldsymbol{\Theta})$. For any $\boldsymbol{W} \in \mathbb{R}^{d\times k}$,*

$$L(\boldsymbol{W}, \mathcal{T}) = s^2 \big\|\boldsymbol{P}_{\mathcal{V}}(\boldsymbol{W} - \boldsymbol{\Theta})\big\|_F^2 + \sigma^2 \big\|\boldsymbol{P}_{\mathcal{V}^\perp}\boldsymbol{W}\big\|_F^2.$$

In particular, the unique minimizer is $\boldsymbol{W}^\star = \boldsymbol{\Theta}$, and $L(\boldsymbol{W}^\star, \mathcal{T}) = 0$.

Proof. Expand the prediction error:

$$\boldsymbol{x}^\top \boldsymbol{W} - \boldsymbol{y} = (\boldsymbol{z} + \boldsymbol{\xi})^\top \boldsymbol{W} - \boldsymbol{z}^\top \boldsymbol{\Theta} = \boldsymbol{z}^\top (\boldsymbol{W} - \boldsymbol{\Theta}) + \boldsymbol{\xi}^\top \boldsymbol{W}.$$

Since $\boldsymbol{z}$ and $\boldsymbol{\xi}$ are independent mean-zero Gaussians, the cross-term vanishes, giving

$$\mathbb{E}\|\boldsymbol{z}^\top (\boldsymbol{W} - \boldsymbol{\Theta}) + \boldsymbol{\xi}^\top \boldsymbol{W}\|_2^2 = \mathbb{E}\|\boldsymbol{z}^\top (\boldsymbol{W} - \boldsymbol{\Theta})\|_2^2 + \mathbb{E}\|\boldsymbol{\xi}^\top \boldsymbol{W}\|_2^2.$$

Using $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = s^2 \boldsymbol{P}_{\mathcal{V}}$ and $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] = \sigma^2 \boldsymbol{P}_{\mathcal{V}^\perp}$,

$$\mathbb{E}\|\boldsymbol{z}^\top (\boldsymbol{W} - \boldsymbol{\Theta})\|_2^2 = \mathrm{Tr}\Big((\boldsymbol{W} - \boldsymbol{\Theta})^\top \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top](\boldsymbol{W} - \boldsymbol{\Theta})\Big) = s^2 \|\boldsymbol{P}_{\mathcal{V}}(\boldsymbol{W} - \boldsymbol{\Theta})\|_F^2,$$
$$\mathbb{E}\|\boldsymbol{\xi}^\top \boldsymbol{W}\|_2^2 = \mathrm{Tr}\Big(\boldsymbol{W}^\top \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top]\boldsymbol{W}\Big) = \sigma^2 \|\boldsymbol{P}_{\mathcal{V}^\perp}\boldsymbol{W}\|_F^2.$$

This proves the formula. Both terms are nonnegative and $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{\Theta} = \boldsymbol{\Theta}$, so the unique minimizer satisfies $\boldsymbol{P}_{\mathcal{V}}\boldsymbol{W} = \boldsymbol{\Theta}$ and $\boldsymbol{P}_{\mathcal{V}^\perp}\boldsymbol{W} = \boldsymbol{0}$, hence $\boldsymbol{W}^\star = \boldsymbol{\Theta}$. □

Lemma D.2 (Eckart–Young Eckart & Young (1936)). *Let $\boldsymbol{M}$ have singular values $\sigma_1(\boldsymbol{M}) \geq \sigma_2(\boldsymbol{M}) \geq \cdots \geq 0$. Then*

$$\min_{\boldsymbol{rank}\boldsymbol{X} \leq p} \|\boldsymbol{M} - \boldsymbol{X}\|_F^2 = \sum_{i>p} \sigma_i(\boldsymbol{M})^2,$$

and a minimizer is the truncated SVD $\boldsymbol{M}^{(p)} = \sum_{i=1}^{p} \sigma_i(\boldsymbol{M})\, \boldsymbol{u}_i \boldsymbol{v}_i^\top$.

Lemma D.3 (Rank splits across orthogonal subspaces). *Let $\boldsymbol{P}$ be an orthogonal projector on $\mathbb{R}^d$ and $\boldsymbol{P}^\perp := \boldsymbol{I} - \boldsymbol{P}$. For any $\boldsymbol{S} \in \mathbb{R}^{d\times k}$, the column spaces of $\boldsymbol{P}\boldsymbol{S}$ and $\boldsymbol{P}^\perp\boldsymbol{S}$ lie in orthogonal subspaces of $\mathbb{R}^d$. In particular,*

$$\boldsymbol{rank}\boldsymbol{S} = \boldsymbol{rank}\boldsymbol{P}\boldsymbol{S} + \boldsymbol{rank}\boldsymbol{P}^\perp\boldsymbol{S}.$$

Proof. $\boldsymbol{P}\boldsymbol{S}$ has columns in $\mathrm{range}(\boldsymbol{P})$ and $\boldsymbol{P}^\perp\boldsymbol{S}$ has columns in $\mathrm{range}(\boldsymbol{P}^\perp)$, which are orthogonal and intersect trivially. Hence $\mathrm{col}(\boldsymbol{S}) = \mathrm{col}(\boldsymbol{P}\boldsymbol{S}) \oplus \mathrm{col}(\boldsymbol{P}^\perp\boldsymbol{S})$ and the rank splits. □

D.2 Exact Characterization of LoRA Minimizers (Rank Allocation)

Lemma D.4 (LoRA finetuning minimizers admit an exact rank-allocation form). *Assume $\mathcal{V}_{\mathrm{base}} \perp \mathcal{V}_{\mathrm{ft}}$, $\boldsymbol{P}_{\mathcal{V}_{\mathrm{ft}}}\boldsymbol{\Theta}_{\mathrm{ft}} = \boldsymbol{\Theta}_{\mathrm{ft}}$, and $\boldsymbol{P}_{\mathcal{V}_{\mathrm{base}}}\boldsymbol{\Theta}_{\mathrm{base}} = \boldsymbol{\Theta}_{\mathrm{base}}$. Let $\boldsymbol{W}_0 = \boldsymbol{\Theta}_{\mathrm{base}}$ and consider*

$$\min_{\boldsymbol{rank}\boldsymbol{S} \leq r} L(\boldsymbol{W}_0 + \boldsymbol{S}, \mathcal{T}_{\mathrm{ft}}).$$

For each $q \in \{0, 1, \ldots, r\}$ define

$$F_r(q) \;:=\; s_{\mathrm{ft}}^2 \sum_{i>q} \sigma_i(\boldsymbol{\Theta}_{\mathrm{ft}})^2 \;+\; \sigma_{\mathrm{ft}}^2 \sum_{i>r-q} \sigma_i(\boldsymbol{\Theta}_{\mathrm{base}})^2, \qquad q^* \in \arg\min_{0\leq q\leq r} F_r(q).$$

Then every global minimizer $\boldsymbol{S}^*$ satisfies

$$\boldsymbol{P}_{\mathcal{V}_{\mathrm{ft}}}(\boldsymbol{W}_0 + \boldsymbol{S}^*) = \boldsymbol{\Theta}_{\mathrm{ft}}^{(q^*)}, \qquad \boldsymbol{P}_{\mathcal{V}_{\mathrm{ft}}^\perp}(\boldsymbol{W}_0 + \boldsymbol{S}^*) = \boldsymbol{\Theta}_{\mathrm{base}} - \boldsymbol{\Theta}_{\mathrm{base}}^{(r-q^*)},$$

and therefore

$$\boldsymbol{W}_{\mathrm{lora}} = \boldsymbol{W}_0 + \boldsymbol{S}^* = \boldsymbol{\Theta}_{\mathrm{ft}}^{(q^*)} + \big(\boldsymbol{\Theta}_{\mathrm{base}} - \boldsymbol{\Theta}_{\mathrm{base}}^{(r-q^*)}\big).$$

Proof. By Lemma D.1 applied to $\mathcal{T}_{\text{ft}}$ and $\mathcal{V}_{\text{base}} \perp \mathcal{V}_{\text{ft}}$ (so $\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}\boldsymbol{\Theta}_{\text{base}} = \mathbf{0}$), for any $\boldsymbol{S}$,

$$\begin{aligned} L(\boldsymbol{W}_0 + \boldsymbol{S}, \mathcal{T}_{\text{ft}}) &= s_{\text{ft}}^2 \big\|\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}(\boldsymbol{W}_0 + \boldsymbol{S} - \boldsymbol{\Theta}_{\text{ft}})\big\|_F^2 + \sigma_{\text{ft}}^2 \big\|\boldsymbol{P}_{\mathcal{V}_{\text{ft}}^\perp}(\boldsymbol{W}_0 + \boldsymbol{S})\big\|_F^2 \\ &= s_{\text{ft}}^2 \big\|\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}\boldsymbol{S} - \boldsymbol{\Theta}_{\text{ft}}\big\|_F^2 + \sigma_{\text{ft}}^2 \big\|\boldsymbol{\Theta}_{\text{base}} + \boldsymbol{P}_{\mathcal{V}_{\text{ft}}^\perp}\boldsymbol{S}\big\|_F^2. \end{aligned} \tag{1}$$

Let $\boldsymbol{X} := \boldsymbol{P}_{\mathcal{V}_{\text{ft}}}\boldsymbol{S}$ and $\boldsymbol{Y} := \boldsymbol{P}_{\mathcal{V}_{\text{ft}}^\perp}\boldsymbol{S}$, so $\boldsymbol{S} = \boldsymbol{X} + \boldsymbol{Y}$. By Lemma D.3, $\mathbf{rank}(\boldsymbol{S}) = \mathbf{rank}(\boldsymbol{X}) + \mathbf{rank}(\boldsymbol{Y})$, hence the constraint $\mathbf{rank}(\boldsymbol{S}) \le r$ is equivalent to $\mathbf{rank}(\boldsymbol{X}) + \mathbf{rank}(\boldsymbol{Y}) \le r$. Thus the problem becomes

$$\min_{\mathbf{rank}(\boldsymbol{X})+\mathbf{rank}(\boldsymbol{Y})\le r} \Big[s_{\text{ft}}^2 \|\boldsymbol{X} - \boldsymbol{\Theta}_{\text{ft}}\|_F^2 + \sigma_{\text{ft}}^2 \|\boldsymbol{\Theta}_{\text{base}} + \boldsymbol{Y}\|_F^2 \Big].$$

Fix $q \in \{0, \dots, r\}$ and impose $\mathbf{rank}(\boldsymbol{X}) \le q$, $\mathbf{rank}(\boldsymbol{Y}) \le r - q$. By Lemma D.2,

$$\min_{\mathbf{rank}(\boldsymbol{X})\le q} \|\boldsymbol{X} - \boldsymbol{\Theta}_{\text{ft}}\|_F^2 = \sum_{i>q} \sigma_i(\boldsymbol{\Theta}_{\text{ft}})^2, \qquad \min_{\mathbf{rank}(\boldsymbol{Y})\le r-q} \|\boldsymbol{\Theta}_{\text{base}} + \boldsymbol{Y}\|_F^2 = \sum_{i>r-q} \sigma_i(\boldsymbol{\Theta}_{\text{base}})^2.$$

Therefore, the optimal value under split q is exactly $F_r(q)$. Minimizing over q yields $q^* \in \arg\min F_r(q)$, and any global minimizer must achieve the Eckart–Young optima at $(q^*, r - q^*)$: $\boldsymbol{X} = \boldsymbol{\Theta}_{\text{ft}}^{(q^*)}$ and $\boldsymbol{Y} = -\boldsymbol{\Theta}_{\text{base}}^{(r-q^*)}$. Substituting back gives the stated form of $\boldsymbol{W}_{\text{lora}}$. □

D.3 Proof of Theorem 3.1

Proof. By Lemma D.1, the unique minimizer of the finetuning population loss is $\boldsymbol{W}_{\text{full}} = \boldsymbol{\Theta}_{\text{ft}}$ and $L(\boldsymbol{W}_{\text{full}}, \mathcal{T}_{\text{ft}}) = 0$.

Evaluate the base loss at $\boldsymbol{W}_{\text{full}} = \boldsymbol{\Theta}_{\text{ft}}$:

$$L(\boldsymbol{\Theta}_{\text{ft}}, \mathcal{T}_{\text{base}}) = s_{\text{base}}^2 \|\boldsymbol{P}_{\mathcal{V}_{\text{base}}}(\boldsymbol{\Theta}_{\text{ft}} - \boldsymbol{\Theta}_{\text{base}})\|_F^2 + \sigma_{\text{base}}^2 \|\boldsymbol{P}_{\mathcal{V}_{\text{base}}^\perp}\boldsymbol{\Theta}_{\text{ft}}\|_F^2.$$

Since $\mathcal{V}_{\text{base}} \perp \mathcal{V}_{\text{ft}}$ and $\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}\boldsymbol{\Theta}_{\text{ft}} = \boldsymbol{\Theta}_{\text{ft}}$, we have $\boldsymbol{P}_{\mathcal{V}_{\text{base}}}\boldsymbol{\Theta}_{\text{ft}} = \mathbf{0}$ and $\boldsymbol{P}_{\mathcal{V}_{\text{base}}^\perp}\boldsymbol{\Theta}_{\text{ft}} = \boldsymbol{\Theta}_{\text{ft}}$, and also $\boldsymbol{P}_{\mathcal{V}_{\text{base}}}\boldsymbol{\Theta}_{\text{base}} = \boldsymbol{\Theta}_{\text{base}}$. Hence

$$L(\boldsymbol{\Theta}_{\text{ft}}, \mathcal{T}_{\text{base}}) = s_{\text{base}}^2 \|\boldsymbol{\Theta}_{\text{base}}\|_F^2 + \sigma_{\text{base}}^2 \|\boldsymbol{\Theta}_{\text{ft}}\|_F^2.$$

Using normalization $s_{\text{base}}^2 \|\boldsymbol{\Theta}_{\text{base}}\|_F^2 = C$ gives $L(\boldsymbol{W}_{\text{full}}, \mathcal{T}_{\text{base}}) \ge C$. □

D.4 Proof of Theorem 3.2

Proof. By Lemma D.1 applied to $\mathcal{T}_{\text{ft}}$, for any $\boldsymbol{W}$,

$$L(\boldsymbol{W}, \mathcal{T}_{\text{ft}}) \;\ge\; s_{\text{ft}}^2 \|\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}(\boldsymbol{W} - \boldsymbol{\Theta}_{\text{ft}})\|_F^2.$$

Under $\mathcal{V}_{\text{base}} \perp \mathcal{V}_{\text{ft}}$ and $\boldsymbol{W}_0 = \boldsymbol{\Theta}_{\text{base}}$, we have $\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}\boldsymbol{W}_0 = \mathbf{0}$. Any LoRA model satisfies $\boldsymbol{W} = \boldsymbol{W}_0 + \boldsymbol{S}$ with $\mathbf{rank}(\boldsymbol{S}) \le r$, hence $\mathbf{rank}(\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}\boldsymbol{W}) = \mathbf{rank}(\boldsymbol{P}_{\mathcal{V}_{\text{ft}}}\boldsymbol{S}) \le r$. Therefore

$$L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{ft}}) \;\ge\; s_{\text{ft}}^2 \min_{\mathbf{rank}(\boldsymbol{X})\le r} \|\boldsymbol{X} - \boldsymbol{\Theta}_{\text{ft}}\|_F^2 = s_{\text{ft}}^2 \sum_{i>r} \sigma_i(\boldsymbol{\Theta}_{\text{ft}})^2,$$

where the equality is Lemma D.2. Next,

$$\sum_{i>r} \sigma_i(\boldsymbol{\Theta}_{\text{ft}})^2 = \|\boldsymbol{\Theta}_{\text{ft}}\|_F^2 - \sum_{i\le r} \sigma_i(\boldsymbol{\Theta}_{\text{ft}})^2 \ge \|\boldsymbol{\Theta}_{\text{ft}}\|_F^2 - r\|\boldsymbol{\Theta}_{\text{ft}}\|_2^2 = \|\boldsymbol{\Theta}_{\text{ft}}\|_F^2 \Big(1 - \frac{r}{\mathbf{intdim}(\boldsymbol{\Theta}_{\text{ft}})}\Big).$$

Truncating at zero and using $s_{\text{ft}}^2 \|\boldsymbol{\Theta}_{\text{ft}}\|_F^2 = C$ yields

$$L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{ft}}) \;\ge\; C\Big(1 - \frac{r}{\mathbf{intdim}(\boldsymbol{\Theta}_{\text{ft}})}\Big)_+.$$

□

D.5 Proof of Theorem 3.3

Proof. Let $r_{\mathrm{ft}} := \mathbf{rank}(\mathbf{\Theta}_{\mathrm{ft}})$ and assume $r \geq r_{\mathrm{ft}}$. We first show that Assumption (A3) implies the rank-allocation optimizer in Lemma D.4 satisfies $q^* = r_{\mathrm{ft}}$. Fix $q < r_{\mathrm{ft}}$ and let $t := r_{\mathrm{ft}} - q \geq 1$. Write $\Delta = r - r_{\mathrm{ft}}$, so $r - q = \Delta + t$. Using the definition of $F_r(\cdot)$ from Lemma D.4,

$$F_r(q) - F_r(r_{\mathrm{ft}}) = s_{\mathrm{ft}}^2 \sum_{i=q+1}^{r_{\mathrm{ft}}} \sigma_i(\mathbf{\Theta}_{\mathrm{ft}})^2 - \sigma_{\mathrm{ft}}^2 \sum_{i=\Delta+1}^{\Delta+t} \sigma_i(\mathbf{\Theta}_{\mathrm{base}})^2$$
$$\geq t\Big(s_{\mathrm{ft}}^2 \sigma_{r_{\mathrm{ft}}}(\mathbf{\Theta}_{\mathrm{ft}})^2 - \sigma_{\mathrm{ft}}^2 \sigma_1(\mathbf{\Theta}_{\mathrm{base}})^2\Big) \;>\; 0,$$

where the final strict inequality is exactly Assumption (A3) after squaring. Hence no $q < r_{\mathrm{ft}}$ can be optimal, and since $r \geq r_{\mathrm{ft}}$, we obtain $q^* = r_{\mathrm{ft}}$.

With $q^* = r_{\mathrm{ft}}$, Lemma D.4 gives

$$\boldsymbol{W}_{\mathrm{lora}} = \mathbf{\Theta}_{\mathrm{ft}} + \big(\mathbf{\Theta}_{\mathrm{base}} - \mathbf{\Theta}_{\mathrm{base}}^{(\Delta)}\big), \qquad \Delta = r - r_{\mathrm{ft}}.$$

Finetuning loss. By Lemma D.1 for $\mathcal{T}_{\mathrm{ft}}$ and $\mathcal{V}_{\mathrm{base}} \perp \mathcal{V}_{\mathrm{ft}}$,

$$L(\boldsymbol{W}_{\mathrm{lora}}, \mathcal{T}_{\mathrm{ft}}) = \sigma_{\mathrm{ft}}^2 \|\mathbf{\Theta}_{\mathrm{base}} - \mathbf{\Theta}_{\mathrm{base}}^{(\Delta)}\|_F^2 \leq \sigma_{\mathrm{ft}}^2 \|\mathbf{\Theta}_{\mathrm{base}}\|_F^2 = \sigma_{\mathrm{ft}}^2 \cdot \frac{C}{s_{\mathrm{base}}^2} = o(1),$$

using normalization and Assumption (A2).

Base loss. By Lemma D.1 for $\mathcal{T}_{\mathrm{base}}$ and $\boldsymbol{P}_{\mathcal{V}_{\mathrm{base}}} \mathbf{\Theta}_{\mathrm{ft}} = \mathbf{0}$,

$$L(\boldsymbol{W}_{\mathrm{lora}}, \mathcal{T}_{\mathrm{base}}) = s_{\mathrm{base}}^2 \|\mathbf{\Theta}_{\mathrm{base}}^{(\Delta)}\|_F^2 + \sigma_{\mathrm{base}}^2 \|\mathbf{\Theta}_{\mathrm{ft}}\|_F^2 = s_{\mathrm{base}}^2 \|\mathbf{\Theta}_{\mathrm{base}}^{(\Delta)}\|_F^2 + o(1),$$

again by normalization and Assumption (A2). Finally,

$$\|\mathbf{\Theta}_{\mathrm{base}}^{(\Delta)}\|_F^2 = \sum_{i \leq \Delta} \sigma_i(\mathbf{\Theta}_{\mathrm{base}})^2 \leq \Delta\, \sigma_1(\mathbf{\Theta}_{\mathrm{base}})^2 = \Delta\, \|\mathbf{\Theta}_{\mathrm{base}}\|_2^2,$$

so

$$L(\boldsymbol{W}_{\mathrm{lora}}, \mathcal{T}_{\mathrm{base}}) \leq s_{\mathrm{base}}^2 \Delta \|\mathbf{\Theta}_{\mathrm{base}}\|_2^2 + o(1) = C \cdot \frac{\Delta}{\mathbf{intdim}(\mathbf{\Theta}_{\mathrm{base}})} + o(1),$$

since $C = s_{\mathrm{base}}^2 \|\mathbf{\Theta}_{\mathrm{base}}\|_F^2$ and $\mathbf{intdim}(\mathbf{\Theta}_{\mathrm{base}}) = \|\mathbf{\Theta}_{\mathrm{base}}\|_F^2 / \|\mathbf{\Theta}_{\mathrm{base}}\|_2^2$. □

D.6 Lower Bounds for the Base Loss

Corollary D.5 (A rank-based lower bound on overshoot cost)**.** *Assume the conditions of Theorem 3.3 and write $\Delta = r - r_{\mathrm{ft}}$. Let $R_{\mathrm{base}} := \boldsymbol{rank}(\mathbf{\Theta}_{\mathrm{base}})$. Then*

$$L(\boldsymbol{W}_{\mathrm{lora}}, \mathcal{T}_{\mathrm{base}}) \;\geq\; C \cdot \frac{\Delta}{R_{\mathrm{base}}} + o(1).$$

Proof. Under Theorem 3.3, we have $L(\boldsymbol{W}_{\mathrm{lora}}, \mathcal{T}_{\mathrm{base}}) = s_{\mathrm{base}}^2 \sum_{i \leq \Delta} \sigma_i(\mathbf{\Theta}_{\mathrm{base}})^2 + o(1)$. For any nonincreasing nonnegative sequence $(a_1, \ldots, a_{R_{\mathrm{base}}})$, $\sum_{i \leq \Delta} a_i \geq \frac{\Delta}{R_{\mathrm{base}}} \sum_{i \leq R_{\mathrm{base}}} a_i$. Applying this to $a_i = \sigma_i(\mathbf{\Theta}_{\mathrm{base}})^2$ yields $\sum_{i \leq \Delta} \sigma_i^2 \geq \frac{\Delta}{R_{\mathrm{base}}} \|\mathbf{\Theta}_{\mathrm{base}}\|_F^2$. Multiplying by s_{base}^2 and using $s_{\mathrm{base}}^2 \|\mathbf{\Theta}_{\mathrm{base}}\|_F^2 = C$ gives the result. □

Assumption D.6 (Flat spectrum for the base task (optional))**.** There exists $\kappa \geq 1$ such that

$$\sigma_i(\mathbf{\Theta}_{\mathrm{base}})^2 \;\geq\; \frac{1}{\kappa}\, \sigma_1(\mathbf{\Theta}_{\mathrm{base}})^2 \qquad \text{for all } 1 \leq i \leq \lfloor \mathbf{intdim}(\mathbf{\Theta}_{\mathrm{base}}) \rfloor.$$

Corollary D.7 (With flat spectrum: a matching **intdim**-based lower bound)**.** *Assume the conditions of Theorem 3.3 and Assumption D.6. Let $\Delta = r - r_{\mathrm{ft}}$ and $d_{\mathrm{base}} := \boldsymbol{intdim}(\mathbf{\Theta}_{\mathrm{base}})$. Then*

$$L(\boldsymbol{W}_{\mathrm{lora}}, \mathcal{T}_{\mathrm{base}}) \;\geq\; \frac{C}{\kappa} \cdot \min\Big\{\frac{\Delta}{d_{\mathrm{base}}}, 1\Big\} + o(1).$$

In particular, for $\Delta \leq d_{\mathrm{base}}$, $L(\boldsymbol{W}_{\mathrm{lora}}, \mathcal{T}_{\mathrm{base}}) \geq \frac{C}{\kappa} \cdot \frac{\Delta}{d_{\mathrm{base}}} + o(1)$, matching the scaling in Theorem 3.3 up to the factor κ.

Proof. Under Theorem 3.3, $L(\boldsymbol{W}_{\text{lora}}, \mathcal{T}_{\text{base}}) = s_{\text{base}}^2 \sum_{i \le \Delta} \sigma_i(\boldsymbol{\Theta}_{\text{base}})^2 + o(1)$. If $\Delta \le \lfloor d_{\text{base}} \rfloor$, Assumption D.6 gives $\sum_{i \le \Delta} \sigma_i^2 \ge \Delta \cdot \sigma_1^2/\kappa$. Using $\sigma_1^2 = \|\boldsymbol{\Theta}_{\text{base}}\|_F^2/d_{\text{base}}$ yields $\sum_{i \le \Delta} \sigma_i^2 \ge (\Delta/(\kappa d_{\text{base}}))\|\boldsymbol{\Theta}_{\text{base}}\|_F^2$. Multiplying by s_{base}^2 and using $s_{\text{base}}^2\|\boldsymbol{\Theta}_{\text{base}}\|_F^2 = C$ gives the desired bound. If $\Delta > \lfloor d_{\text{base}} \rfloor$, then $\sum_{i \le \Delta} \sigma_i^2 \ge \sum_{i \le \lfloor d_{\text{base}} \rfloor} \sigma_i^2 \ge \lfloor d_{\text{base}} \rfloor \sigma_1^2/\kappa \gtrsim \|\boldsymbol{\Theta}_{\text{base}}\|_F^2/\kappa$, which yields the saturation at $\gtrsim C/\kappa$ (absorbing flooring into constants and $o(1)$). □

D.7 Discussion on Gradient Descent

Recall the population loss on a task $\mathcal{T} = (\mathcal{V}, s, \sigma, \boldsymbol{\Theta})$:

$$L(\boldsymbol{W}, \mathcal{T}) = \mathbb{E}\big\|\boldsymbol{x}^\top \boldsymbol{W} - \boldsymbol{y}\big\|_2^2, \qquad \boldsymbol{x} = \boldsymbol{z} + \boldsymbol{\xi},\ \boldsymbol{z} \sim \mathcal{N}(0, s^2 \boldsymbol{P}_{\mathcal{V}}),\ \boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{P}_{\mathcal{V}^\perp}), \qquad \boldsymbol{y} = \boldsymbol{z}^\top \boldsymbol{\Theta}.$$

Let $\boldsymbol{\Sigma}_x := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = s^2 \boldsymbol{P}_{\mathcal{V}} + \sigma^2 \boldsymbol{P}_{\mathcal{V}^\perp}$ and $\boldsymbol{\Sigma}_{xy} := \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] = s^2 \boldsymbol{P}_{\mathcal{V}} \boldsymbol{\Theta}$. Then

$$L(\boldsymbol{W}, \mathcal{T}) = \operatorname{Tr}(\boldsymbol{W}^\top \boldsymbol{\Sigma}_x \boldsymbol{W}) - 2 \operatorname{Tr}(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{xy}) + \operatorname{Tr}(\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]). \tag{2}$$

By Lemma D.1, the minimizer of $L(\boldsymbol{W}, \mathcal{T})$ is unique and is given by $\boldsymbol{W}^\star = \boldsymbol{\Theta}$.

Full-model finetuning. Consider the finetuning task $\mathcal{T}_{\text{ft}} = (\mathcal{V}_{\text{ft}}, s_{\text{ft}}, \sigma_{\text{ft}}, \boldsymbol{\Theta}_{\text{ft}})$ with $s_{\text{ft}} > 0$ and $\sigma_{\text{ft}} > 0$. We run gradient descent on $\boldsymbol{W}$:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla_{\boldsymbol{W}} L(\boldsymbol{W}_t, \mathcal{T}_{\text{ft}}) = \boldsymbol{W}_t - 2\eta\big(\boldsymbol{\Sigma}_x^{\text{ft}} \boldsymbol{W}_t - \boldsymbol{\Sigma}_{xy}^{\text{ft}}\big).$$

Let $\lambda_{\max} := \lambda_{\max}(\boldsymbol{\Sigma}_x^{\text{ft}}) = \max\{s_{\text{ft}}^2, \sigma_{\text{ft}}^2\}$. If $0 < \eta < 1/\lambda_{\max}$, then gradient descent converges linearly to the unique minimizer $\boldsymbol{W}_{\text{full}} = \boldsymbol{\Theta}_{\text{ft}}$:

$$\boldsymbol{W}_t - \boldsymbol{W}_{\text{full}} = (\boldsymbol{I} - 2\eta\, \boldsymbol{\Sigma}_x^{\text{ft}})^t (\boldsymbol{W}_0 - \boldsymbol{W}_{\text{full}}),$$

and hence $\boldsymbol{W}_t \to \boldsymbol{W}_{\text{full}}$ as $t \to \infty$.

LoRA finetuning. In the population setting, the finetuning loss is a quadratic function of $\boldsymbol{W}$:

$$L(\boldsymbol{W}, \mathcal{T}_{\text{ft}}) = \operatorname{Tr}(\boldsymbol{W}^\top \boldsymbol{\Sigma}_x^{\text{ft}} \boldsymbol{W}) - 2 \operatorname{Tr}(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{xy}^{\text{ft}}) + \operatorname{Tr}(\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]),$$

where $\boldsymbol{\Sigma}_x^{\text{ft}} \succ 0$ since $s_{\text{ft}}, \sigma_{\text{ft}} > 0$. Completing the square gives

$$L(\boldsymbol{W}, \mathcal{T}_{\text{ft}}) = \big\|(\boldsymbol{\Sigma}_x^{\text{ft}})^{1/2} \boldsymbol{W} - (\boldsymbol{\Sigma}_x^{\text{ft}})^{-1/2} \boldsymbol{\Sigma}_{xy}^{\text{ft}}\big\|_F^2 + \text{const}.$$

With the LoRA parameterization $\boldsymbol{W} = \boldsymbol{W}_0 + \boldsymbol{B}\boldsymbol{A}$, define

$$\boldsymbol{M} := (\boldsymbol{\Sigma}_x^{\text{ft}})^{-1/2} \boldsymbol{\Sigma}_{xy}^{\text{ft}} - (\boldsymbol{\Sigma}_x^{\text{ft}})^{1/2} \boldsymbol{W}_0.$$

Then minimizing $L(\boldsymbol{W}_0 + \boldsymbol{B}\boldsymbol{A}, \mathcal{T}_{\text{ft}})$ over $(\boldsymbol{B}, \boldsymbol{A})$ is equivalent (up to an additive constant and an invertible linear change of variables in $\boldsymbol{B}$) to the classical matrix factorization / shallow linear network objective

$$\min_{\boldsymbol{B} \in \mathbb{R}^{d \times r},\, \boldsymbol{A} \in \mathbb{R}^{r \times k}} \|(\boldsymbol{\Sigma}_x^{\text{ft}})^{1/2} \boldsymbol{B}\boldsymbol{A} - \boldsymbol{M}\|_F^2.$$

Classical results in shallow linear networks / matrix factorization imply that this objective has a benign landscape. To see this, let $\boldsymbol{C} := (\boldsymbol{\Sigma}_x^{\text{ft}})^{1/2} \succ 0$ and reparameterize $\boldsymbol{U} := \boldsymbol{C}\boldsymbol{B}$. Since $\boldsymbol{C}$ is invertible, the map $(\boldsymbol{B}, \boldsymbol{A}) \mapsto (\boldsymbol{U}, \boldsymbol{A})$ is a bijection and

$$\min_{\boldsymbol{B}, \boldsymbol{A}} \|\boldsymbol{C}\boldsymbol{B}\boldsymbol{A} - \boldsymbol{M}\|_F^2 \quad \equiv \quad \min_{\boldsymbol{U}, \boldsymbol{A}} \|\boldsymbol{U}\boldsymbol{A} - \boldsymbol{M}\|_F^2,$$

which is the classical matrix factorization / shallow linear network objective. Moreover, invertible linear changes of variables preserve critical points and the presence of negative curvature (i.e., $(\boldsymbol{B}, \boldsymbol{A})$ is critical iff $(\boldsymbol{U}, \boldsymbol{A})$ is critical, and the Hessian has a negative eigenvalue at one iff it does at the other). For this objective, every local minimum is a global minimum, and every critical point that is not globally optimal is a saddle point (Baldi & Hornik, 1989; Kawaguchi, 2016; Valavi et al., 2020; Zhu et al., 2021). Consequently, to argue global optimality of gradient descent, it suffices to show that it converges to a local minimum rather than a saddle. Under the strict-saddle geometry, gradient descent with random initialization (almost surely) does not converge to strict saddles and instead converges to a local minimizer (Lee et al., 2016; Panageas & Piliouras, 2016). Together with the absence of spurious local minima, this implies that gradient descent on the LoRA parameters converges to a global minimizer under mild conditions.