OASIS: Mitigating Harmful Fine-tuning Attacks on LLMs via Orthogonal and Adaptive Safety Alignment Strategy

Jiayu Tang[1], Guowei Peng[1], Qiuhao Xie[1], Yuning Yang[1], Xiurui Xie[1]*, Guisong Liu[2]*
[1]University of Electronic Science and Technology of China, Chengdu, China
[2]Southwestern University of Finance and Economics, Chengdu, China
{tangjiayu, pengguowei, qiuhao.xie, yangyuning}@std.uestc.edu.cn
xiexiurui@uestc.edu.cn, gliu@swufe.edu.cn

Abstract

The "Fine-Tuning-as-a-Service" paradigm exposes large language models to catastrophic safety degradation from less harmful samples. Alignment-stage defenses address this by proactively injecting adversarial perturbations to bolster the model's inherent robustness against harmful drift. However, existing methods rely on perturbation directions that often conflict with harmful gradients, inadvertently facilitating the acquisition of malicious features rather than suppressing them. To address this issue, we propose **O**rthogonal and **A**daptive **S**afety **A**lignment **S**trategy **(OASIS)** to mathematically decouple safety enforcement from harmful feature acquisition. By projecting perturbations orthogonal to harmful gradients and concentrating optimization on adaptively selected safety-critical layers, OASIS effectively resolves directional conflicts while maximizing parameter efficiency. Extensive experiments on four LLMs across three datasets (SST2, GSM8K, and AGNews) demonstrate that OASIS reduces the Harmful Score by approximately **60%** compared to competitive baselines, while maintaining stable downstream task utility. Our code is publicly available at https://github.com/xiaoroyi/OASIS.

1 Introduction

With the rapid advancement of Large Language Models (LLMs), "Fine-Tuning-as-a-Service" (FTaaS) has emerged as a dominant business model, enabling users to adapt generic models to domain-specific scenarios via custom data. However, this accessibility presents a severe security threat: *harmful fine-tuning attacks* (Qi et al., 2023; Yang et al., 2023). Empirical research demonstrates that attackers can severely compromise a model's safety alignment by injecting a minimal number of harmful samples, inducing the generation of inappropriate or dangerous content. This vulnerability poses a significant challenge for service providers aiming to uphold rigorous safety standards. To mitigate such risks, defense strategies have been proposed at the alignment, fine-tuning, and post-fine-tuning stages. Among these, alignment-stage defense is recognized as the most resource-efficient paradigm, as it incurs a one-time computational cost for the service provider, avoiding the cumulative overhead required by defenses that intervene during individual user fine-tuning session (Huang et al., 2024d).

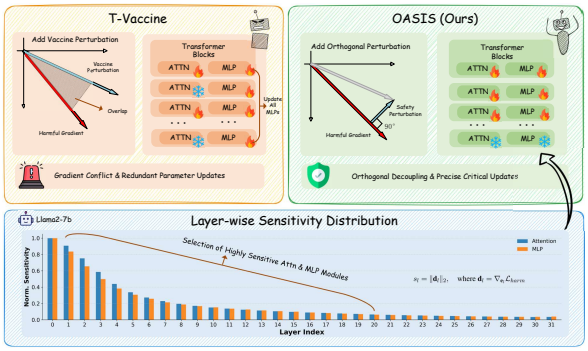

Figure 1: **Motivation and Method Overview.** (Top) Comparison of defense mechanisms. (Bottom) Layer-wise sensitivity distribution. The sensitivity score s_l is defined as the L_2-norm of the harmful gradient w.r.t. the Attention and MLP modules respectively.

Among representative alignment-stage defenses, Vaccine (Huang et al., 2024d) introduces a perturbation-aware alignment strategy to counteract harmful embedding drift, applying global gradient-guided noise to bolster model robustness. Subsequently, targeting the issue of resource efficiency, T-Vaccine (Liu et al., 2025) proposes a lightweight alternative. It employs a specific heuristic that restricts perturbations to only 8 attention layers while updating all MLP modules to reduce memory overhead.

Although Vaccine and T-Vaccine bolster robustness, we identify two critical limitations reducing their efficacy. First, regarding perturbation *direc-*

*Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28407–28421
July 2-7, 2026 ©2026 Association for Computational Linguistics

tion, our analysis reveals a critical Directional Overlap. As illustrated in Figure 1 (top), baseline perturbations often explicitly overlap with harmful gradients, inadvertently facilitating malicious feature acquisition. Second, regarding defense *scope*, existing methods are not sufficiently targeted. As evidenced by the sensitivity distribution in Figure 1 (bottom), model vulnerability exhibits a distinct layer-wise decaying trend, rendering uniform update strategies resource-inefficient.

To address these challenges, we propose **OASIS** to simultaneously resolve directional conflicts and optimize resource allocation. First, we introduce an **orthogonal perturbation** mechanism. By mathematically projecting safety perturbations onto the subspace orthogonal to the harmful gradient, we decouple safety enforcement from harmful feature acquisition, ensuring updates occur strictly within a "safe subspace". Second, we implement an **adaptive safety module selection strategy**. Driven by the observation that sensitivity is primarily layer-dependent, we identify and select the **top-K safety-critical layers**. Within these layers, we synchronously update both Attention and MLP modules to maximize parameter efficiency.

Our main contributions are as follows:

- We identify the phenomenon of *gradient conflict* in existing perturbation-based defenses and propose an **orthogonal perturbation mechanism** to eliminate the directional overlap between safety alignment and harmful gradients.
- We propose an **adaptive safety module selection strategy** based on dynamic sensitivity analysis. By targeting the most vulnerable layers and synchronously updating their internal modules, we achieve superior defense efficiency under a strict parameter budget.
- Comprehensive evaluations across multiple LLM architectures (Llama2, Qwen2, Gemma2, Vicuna) and downstream tasks (SST2, GSM8K, and AGNews) demonstrate that OASIS significantly outperforms baselines, achieving the lowest harmful score while maintaining superior downstream task accuracy.

2 Related Work

Alignment-Stage Defenses. Alignment-stage defenses proactively immunize models against safety degradation prior to fine-tuning. A dominant paradigm is *perturbation-aware alignment*, pioneered by Vaccine (Huang et al., 2024d), which injects worst-case invariant perturbations into hidden embeddings to mitigate harmful drift. To reduce computational overhead, T-Vaccine (Liu et al., 2025) optimizes this approach via a layer-wise strategy, targeting perturbations at safety-critical modules. Complementing these gradient-based methods, Booster (Huang et al., 2024b) explicitly attenuates harmful loss reduction. Beyond perturbation, researchers explore precise representation control: RepNoise (Rosati et al., 2024) minimizes the mutual information between harmful representations and inputs, while RSN-Tune (Zhao et al., 2025) selectively tunes specific safety neurons at a finer granularity. Furthermore, active defenses like TAR (Tamirisa et al., 2024) (adversarial training) and SDD (Chen et al., 2025) (self-degradation) poison the attacker's objective. Relatedly, BackdoorAlign (Wang et al., 2024) mitigates jailbreaks via secret data-level triggers, complementing parameter-level approaches like ours.

Fine-Tuning-Stage Defenses. Fine-tuning-stage defenses intervene during the user's adaptation process, primarily focusing on data curation and optimization constraints. In data curation, approaches range from bilevel optimization in SEAL (Shen et al., 2024) and dataset mixing in VLGuard (Zong et al., 2024), to dynamic Bayesian scheduling in BDS (Hu et al., 2025) for precise weighting. Parallel to this, parameter-level interventions like Lisa (Huang et al., 2024c) and Constrain-SFT (Qi et al., 2024) apply regularization terms to restrict deviations from the aligned anchor. More recently, gradient surgery techniques like SafeGrad (Yi et al., 2025) and SaLoRA (Li et al., 2025) advance this direction by projecting updates into safety-orthogonal subspaces to reconcile plasticity with stability.

Post-Fine-Tuning Defenses. As remedial measures, these strategies recover safety mechanisms after the model is compromised. Post-hoc unlearning techniques, such as Antidote (Huang et al., 2024a) and LAT (Casper et al., 2024), are complemented by Eraser (Lu et al., 2024), which precisely identifies and prunes harmful knowledge parameters. For inference-time defense, Panacea (Wang et al., 2025) and SafetyLock (Zhu et al., 2024) utilize adaptive noise injection and parameter locking to disrupt harmful generation. Furthermore, subspace fusion methods like SOMF (Yi et al.,

2024) progress toward selective merging frameworks like SafeMERGE (Djuhera et al., 2025), to efficiently reconstruct safety barriers from compromised weights.

3 Methodology

3.1 Problem Formulation

We focus on the standard FTaaS lifecycle, which involves two distinct stages:

Safety Alignment. The service provider optimizes a pre-trained base model $M_{\boldsymbol{\theta}}$ using a high-quality, sanitized dataset $\mathcal{D}_{align} = \{(x, y)\}$. The objective is to obtain aligned parameters $\boldsymbol{\theta}_{aligned}$ that minimize the safety alignment loss $\mathcal{L}_{align}$, ensuring the model follows instructions safely.

User Fine-Tuning (Threat Model). Subsequently, users fine-tune the aligned model on a custom dataset $\mathcal{D}_{user}$. We consider a realistic threat model where an attacker poisons $\mathcal{D}_{user}$ by injecting a small proportion of harmful instructions (e.g., malicious queries). The attacker's goal is to leverage the plasticity of the fine-tuning process to degrade the model's safety barriers, creating a "jailbroken" model that generates harmful content for unseen malicious queries.

3.2 Revisiting Perturbation-Aware Alignment and Gradient Conflict

Representative defenses, exemplified by Vaccine (Huang et al., 2024d), mitigate safety degradation through the lens of *Robust Optimization*. Rather than solely minimizing the standard empirical risk on clean data, these methods seek to minimize the worst-case loss within a constrained neighborhood of the latent representations, thereby enforcing stability and smoothness in the optimization landscape.

Formally, consider an input embedding vector $\mathbf{e} \in \mathbb{R}^d$ corresponding to a safety alignment sample (x, y). The robust alignment objective is formulated as a min-max problem:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{align}} \left[\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}_{align}(f(\mathbf{e} + \boldsymbol{\epsilon}); y) \right] \tag{1}$$

where ρ denotes the perturbation radius constraints. To solve the inner maximization, existing approaches typically employ a first-order Taylor expansion to approximate the loss landscape linearly. Under this approximation, the optimal adversarial perturbation $\boldsymbol{\epsilon}^*_{std}$ admits a closed-form solution aligned with the gradient direction:

$$\boldsymbol{\epsilon}^*_{std} \approx \rho \frac{\nabla_{\mathbf{e}} \mathcal{L}_{align}}{\|\nabla_{\mathbf{e}} \mathcal{L}_{align}\|_2} \tag{2}$$

This standard perturbation directs the model to effectively suppress sensitivity along the direction of steepest ascent for the alignment loss. While effective for general model robustness, we postulate that this directionality, which remains constrained strictly parallel to the safety alignment gradient, is suboptimal for defending against targeted harmful fine-tuning scenarios.

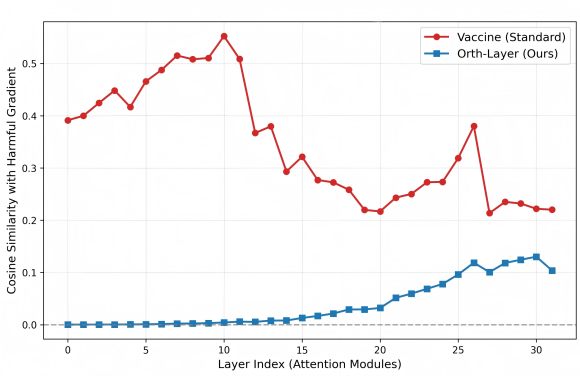

Figure 2: **Gradient Conflict Analysis.** Comparison of the cosine similarity between the safety alignment gradient and the harmful gradient across layers for Vaccine (Red) and OASIS (Blue).

To validate this hypothesis, we conduct an empirical investigation into the geometric relationship between the *safety gradient* ($\mathbf{g} = \nabla_{\mathbf{e}} \mathcal{L}_{align}$) and the *harmful gradient* ($\mathbf{d} = \nabla_{\mathbf{e}} \mathcal{L}_{harm}$). Specifically, we calculate the gradients with respect to the hidden states (activations) of each layer on the Llama2-7B. We then quantify their directional overlap by computing the cosine similarity $\cos(\mathbf{g}, \mathbf{d})$, averaged over 1,000 sampled instances from the alignment and harmful datasets.

As illustrated in Figure 2, the baseline approach (Vaccine, depicted in red) exhibits a consistently positive cosine similarity between $\mathbf{g}$ and $\mathbf{d}$ across model layers, with values typically ranging from 0.2 to 0.5. This persistent positive correlation indicates a significant structural overlap between the gradient subspaces of safety alignment and harmful fine-tuning.

This geometric alignment exposes a fundamental vulnerability in existing defenses. Given that the standard perturbation is constructed as $\boldsymbol{\epsilon}_{std} \propto \mathbf{g}$, the condition $\mathbf{g}^\top \mathbf{d} > 0$ implies:

$$\boldsymbol{\epsilon}^\top_{std} \mathbf{d} > 0 \tag{3}$$

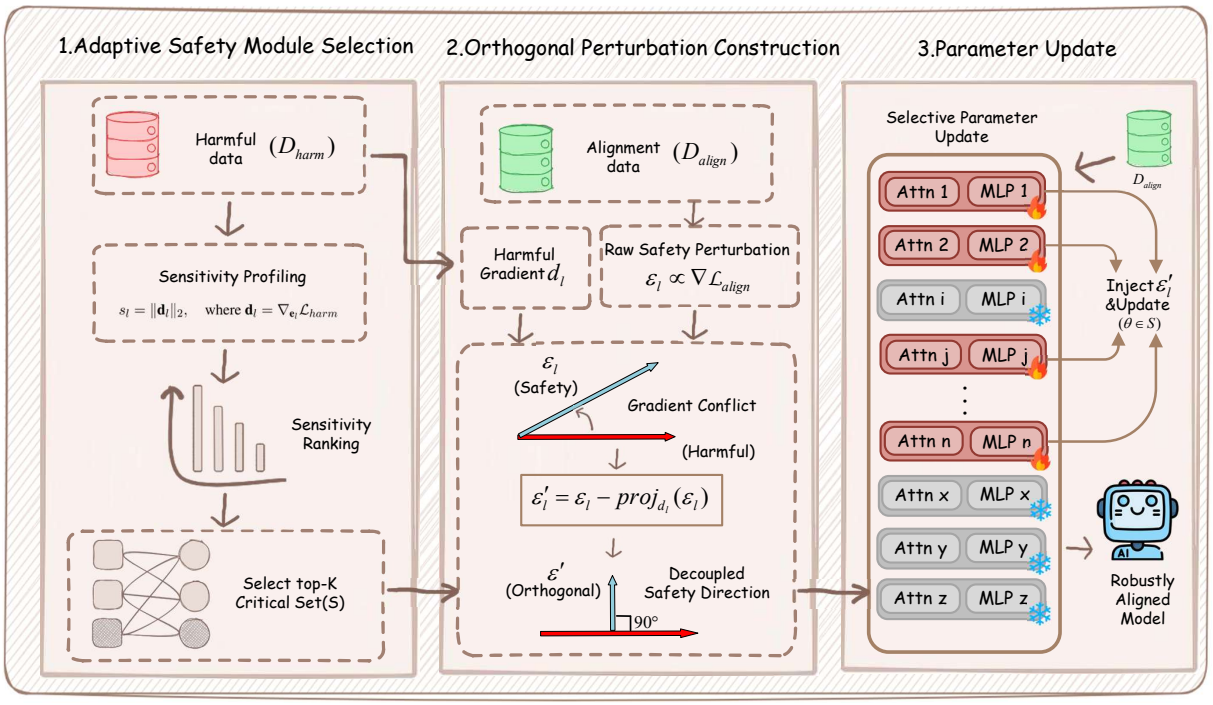

Figure 3: **Overview of the OASIS framework.** The pipeline comprises three stages: (1) **Adaptive Safety Module Selection** identifies the top-K critical layers ($\mathcal{S}$) via sensitivity profiling; (2) **Orthogonal Perturbation Construction** projects the raw safety perturbation onto the subspace orthogonal to the harmful gradient to resolve gradient conflicts; (3) **Selective Parameter Update** synchronously optimizes the Attention and MLP parameters of the selected layers using the robust loss derived from the noise-injected hidden states.

This inequality indicates that the standard perturbation projects positively onto the harmful gradient. Consequently, optimizing along $\boldsymbol{\epsilon}_{std}$ implicitly involves a directional component overlapping with the attacker's objective. We term this phenomenon *Gradient Conflict*, which motivates OASIS to explicitly enforce orthogonality and decouple safety from harmful feature acquisition.

Based on the analysis of gradient conflict, we propose **OASIS**, a unified framework designed to decouple safety optimization from harmful subspaces. As illustrated in Figure 3, our approach synergizes Adaptive Safety-Sensitive Module Selection to optimize the defense scope and Orthogonal Perturbation Construction to resolve directional interference, culminating in a Selective Parameter Update strategy for efficient alignment.

3.3 Adaptive Safety-Sensitive Module Selection

Safety-critical information within large language models is non-uniformly distributed across layers rather than being ubiquitous. Consequently, indiscriminate global updates are computationally inefficient and risk degrading general capabilities by modifying unrelated benign features. Therefore, our first objective is to deterministically identify the precise subset of parameters most susceptible to harmful instructions.

We introduce a metric termed *Harmful Sensitivity Score* (s_l). Formally, given a harmful dataset $\mathcal{D}_{harm}$ and the loss function $\mathcal{L}_{harm}$, the sensitivity of the l-th layer is quantified by the L_2-norm of its gradient with respect to the input embeddings $\mathbf{e}_l$:

$$s_l = \|\mathbf{d}_l\|_2, \quad \text{where } \mathbf{d}_l = \nabla_{\mathbf{e}_l} \mathcal{L}_{harm}(\theta; \mathcal{D}_{harm}). \tag{4}$$

Intuitively, s_l measures the magnitude of activation change required to minimize the harmful loss, serving as a proxy for the layer's "willingness" to accept harmful knowledge.

Selection Strategy. As visualized in Figure 1 (bottom), sensitivity is primarily layer-dependent, exhibiting a significant decay with depth. Since the variance between layers outweighs the discrepancy between Attention and MLP modules, we simplify selection to layer level. We calculate the aggregated sensitivity for each layer and select the top-K layers to form the critical set $\mathcal{S}$. Within these layers, we synchronously update Attention and MLP modules. To mitigate computational overhead, $\mathcal{S}$ is updated periodically rather than at every iteration.

3.4 Orthogonal Perturbation Construction

Standard adversarial training typically generates perturbations that maximize the alignment loss to improve robustness. However, we argue that without geometric constraints, such perturbations can inadvertently drift towards the harmful subspace.

Gradient Conflict Analysis. Let $\mathcal{L}_{align}$ denote the alignment loss. In standard robust alignment, a perturbation $\boldsymbol{\epsilon}$ is crafted to maximize $\mathcal{L}_{align}$, typically following the direction of the gradient $\mathbf{g}_l = \nabla_{\mathbf{e}_l} \mathcal{L}_{align}$. A critical issue arises when the alignment gradient and the harmful gradient are positively correlated. Consider the first-order Taylor expansion of the change in harmful loss induced by the perturbation $\boldsymbol{\epsilon}$:

$$\Delta \mathcal{L}_{harm} \approx \boldsymbol{\epsilon}^\top \nabla_{\mathbf{e}_l} \mathcal{L}_{harm} = \boldsymbol{\epsilon}^\top \mathbf{d}_l. \quad (5)$$

If we set $\boldsymbol{\epsilon} \propto \mathbf{g}_l$ (as in standard methods), the change becomes proportional to the inner product $\langle \mathbf{g}_l, \mathbf{d}_l \rangle$. Empirical evidence suggests that $\cos(\mathbf{g}_l, \mathbf{d}_l)$ is often non-zero and positive in critical layers (gradient conflict), implying that maximizing alignment robustness implicitly increases the model's susceptibility to harmful instructions ($\Delta \mathcal{L}_{harm} > 0$).

Orthogonal Projection. To resolve this dilemma, we formally derive our perturbation by solving a constrained maximization problem. We seek the optimal perturbation direction $\boldsymbol{\epsilon}^*$ that maximizes the alignment gain within the local trust region, subject to a strict orthogonality constraint against the harmful gradient, thereby mathematically decoupling the safety and harmful subspaces:

$$\max_{\boldsymbol{\epsilon}} \ \mathbf{g}_l^\top \boldsymbol{\epsilon} \quad \text{s.t.} \quad \boldsymbol{\epsilon} \perp \mathbf{d}_l, \quad \|\boldsymbol{\epsilon}\|_2 \leq \rho. \quad (6)$$

To enforce the safety constraint, we derive the final perturbation by projecting the standard gradient-based perturbation onto the orthogonal subspace of the harmful gradient. Let $\boldsymbol{\epsilon}_l$ be the standard perturbation normalized to the trust region ρ (i.e., $\|\boldsymbol{\epsilon}_l\| = \rho$). The *orthogonal perturbation* $\boldsymbol{\epsilon}'_l$ is obtained by subtracting its projection on $\mathbf{d}_l$:

$$\boldsymbol{\epsilon}'_l = \boldsymbol{\epsilon}_l - \text{proj}_{\mathbf{d}_l}(\boldsymbol{\epsilon}_l) = \boldsymbol{\epsilon}_l - \frac{\boldsymbol{\epsilon}_l^\top \mathbf{d}_l}{\|\mathbf{d}_l\|^2 + \gamma} \mathbf{d}_l, \quad (7)$$

where γ is a smoothing term. Since projection is a non-expansive operator, $\|\boldsymbol{\epsilon}'_l\| \leq \|\boldsymbol{\epsilon}_l\| = \rho$, ensuring the update remains strictly within the defined trust region.

3.5 Selective Parameter Update

With the critical set $\mathcal{S}$ identified via the gradient mapping and the orthogonal perturbation $\boldsymbol{\epsilon}'$ constructed, we proceed to the parameter update phase. Unlike methods that update the entire model, we employ a selective update rule to concentrate the defense budget strictly on the most vulnerable regions identified by our sensitivity analysis.

The training objective is to minimize the robust alignment loss under the orthogonal perturbation. The final loss function is defined as:

$$\mathcal{L}_{total} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_{align}} \left[\mathcal{L}_{align} \left(f(\mathbf{x} + \boldsymbol{\epsilon}'; \theta) \right) \right], \quad (8)$$

where the perturbation $\boldsymbol{\epsilon}'$ is only injected into the layers belonging to $\mathcal{S}$. Accordingly, the parameters θ are updated via gradient descent:

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla_\theta \mathcal{L}_{total} \cdot \mathbb{I}(\theta \in \mathcal{S}), \quad (9)$$

where $\mathbb{I}(\cdot)$ is the indicator function. This strategy ensures that only the modules identified as safety-sensitive (in Section 3.3) are modified, while the remaining parameters are frozen to maintain the model's original utility. The overall training procedure is summarized in Algorithm 1.

Algorithm 1 OASIS Training Procedure

Input: Model θ, Alignment Data $\mathcal{D}_{align}$, Harmful Data $\mathcal{D}_{harm}$.
Hyperparams: Perturbation radius ρ, learning rate η, Update Frequency T_{freq}.
Initialize critical set $\mathcal{S} \leftarrow \emptyset$, harmful gradients $\mathbf{d} \leftarrow \emptyset$.
for step $t = 1, \ldots, T$ **do** ▷ Iterate over training steps
  Get minibatch $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{align}$.
  if $t \pmod{T_{freq}} == 0$ **then**
    // Stage 1: Adaptive Selection (Periodic)
    Sample $(\mathbf{x}_{harm}, \mathbf{y}_{harm}) \in \mathcal{D}_{harm}$.
    Compute $\mathbf{d}_l \leftarrow \nabla_{\mathbf{e}_l} \mathcal{L}_{harm}$ for all layers.
    Update sensitivity $s_l \leftarrow \|\mathbf{d}_l\|_2$.
    Update $\mathcal{S} \leftarrow$ Top-$K(s_l)$.
  end if
  // Stage 2: Orthogonal Perturbation Construction
  Compute alignment gradient $\mathbf{g}_l$ on $(\mathbf{x}, \mathbf{y})$.
  Generate perturbation $\boldsymbol{\epsilon}_l \leftarrow \rho \cdot \mathbf{g}_l / \|\mathbf{g}_l\|_2$. ▷ Standard Normalization
  for $l \in \mathcal{S}$ **do**
    Retrieve cached $\mathbf{d}_l$.
    Project: $\boldsymbol{\epsilon}'_l \leftarrow \boldsymbol{\epsilon}_l - \text{proj}_{\mathbf{d}_l}(\boldsymbol{\epsilon}_l)$ (Eq. 7).
  end for
  // Stage 3: Selective Parameter Update
  Compute robust loss $\mathcal{L}_{total}$ with injected $\boldsymbol{\epsilon}'$.
  Update θ: $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{total} \cdot \mathbb{I}(\theta \in \mathcal{S})$.
end for

Table 1: Main results on Llama2-7B for the **SST2 task**. We highlight the **lowest** Harmful Score (HS) and **competitive** Finetune Accuracy (AC) among defense methods.

Method	Harmful Score (HS) ↓					Finetune Accuracy (AC) ↑					Parameters ↓
	$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	Average	$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	Average	
No-Align	79.4	79.8	82.2	81.0	80.6	93.4	93.6	93.2	93.0	93.3	-
Vaccine	6.2	20.4	41.2	60.2	32.0	91.0	91.2	90.8	90.8	91.0	19.9M
T-Vaccine	8.0	16.2	35.4	51.8	27.9	91.0	91.8	91.2	90.6	91.2	13.6M
Booster	28.0	27.8	31.8	37.0	31.5	94.4	94.4	94.2	93.2	94.1	19.9M
LISA	6.8	12.0	28.6	41.0	22.1	94.4	93.8	92.8	92.8	93.5	19.9M
SafeLoRA	36.6	62.4	76.6	80.2	64.0	93.4	93.6	93.6	92.8	93.4	-
OASIS	**1.6**	**5.4**	**13.6**	**36.8**	**14.4**	92.0	92.2	91.6	90.2	91.5	**12.4M**

4 Experiments

4.1 Experimental Setup

Datasets. Our experimental pipeline consists of two stages: the alignment stage and the user fine-tuning stage. In the *alignment phase*, we randomly sample the BeaverTails dataset (Ji et al., 2023) to construct a safety alignment dataset, consisting of 2,000 safe samples and 200 harmful samples. In the *user fine-tuning phase*, we evaluate the model performance on three downstream task datasets: SST2 (sentiment analysis) (Socher et al., 2013), GSM8K (mathematical reasoning) (Cobbe et al., 2021), and AGNews (text classification) (Zhang et al., 2015). To simulate harmful fine-tuning attack, we mix harmful samples from BeaverTails into the downstream task datasets at different contamination ratios $p \in 0.05, 0.1, 0.2, 0.4$. In all experimental settings, the user fine-tuning dataset size is fixed at 1,000 samples.

Models. We conduct our primary evaluation on Llama2-7B (Touvron et al., 2023). To assess architectural universality, we extend experiments to Qwen2-7B (Yang et al., 2024), Gemma2-9B (Team et al., 2024), and Vicuna-7B (Zheng et al., 2023).

Baselines. To comprehensively evaluate our method, we compare OASIS against representative defenses across all three intervention stages: **No-Align** (Standard SFT without prior safety alignment); Alignment-Stage Defenses including **Vaccine** (Global perturbation) (Huang et al., 2024d), **T-Vaccine** (Targeted Attention perturbation) (Liu et al., 2025), and **Booster** (Harmful loss attenuation) (Huang et al., 2024b); Fine-Tuning-Stage Defense including **LISA** (Huang et al., 2024c); and Post-Fine-Tuning Defense including **SafeLoRA** (Hsu et al., 2024).

Implementation Details. All models are trained using LoRA (Hu et al., 2021) on NVIDIA A800 GPUs. For fair comparison, Booster and LISA are instantiated with the same LoRA backbone as OASIS, Vaccine, and T-Vaccine (i.e., rank-8 adapters on standard transformer projection modules), while SafeLoRA follows its original post-fine-tuning projection pipeline. For detailed hyperparameters, optimization settings, and model-specific configurations, please refer to **Appendix B**.

Metrics. We adopt two key metrics: **Harmful Score (HS ↓):** We utilize a robust moderation model (Ji et al., 2023) to classify the model's responses to unseen malicious instructions. HS measures the percentage of outputs classified as unsafe. **Finetune Accuracy (AC ↑):** We report the standard accuracy on the respective test sets to evaluate the preservation of model utility.

4.2 Main Results

Defense Efficacy and Utility Preservation. We evaluate the defense performance on Llama2-7B across three distinct downstream tasks. As presented in Table 1, to firmly establish the State-of-the-Art (SOTA), we comprehensively compare OASIS against representative methods from all three defense stages on the SST2 dataset. OASIS demonstrates superior robustness and utility, achieving a comprehensive reduction in Harmful Score (HS) across all four attack intensities. Notably, OASIS consistently achieves the lowest average HS (14.4), significantly outperforming the during-finetuning defense LISA (22.1) and the alignment-stage defense Booster (31.5). This demonstrates that despite being an alignment-stage defense—which operates proactively prior to the attacker's intervention—OASIS maintains SOTA performance even when compared with methods that involve strong interventions during the actual fine-tuning process. Furthermore, our method preserves highly competitive task accuracy (91.5% on average). In terms of

Table 2: Main results on Llama2-7B for **GSM8K and AGNews** tasks. We highlight the **lowest** Harmful Score (HS) and **competitive** Finetune Accuracy (AC) among defense methods.

Task	Method	Harmful Score (HS) ↓					Finetune Accuracy (AC) ↑					Parameters ↓
		$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	Average	$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	Average	
GSM8K	No-Align	80.4	81.6	80.8	80.8	80.9	19.2	18.8	17.6	18.2	18.5	-
	Vaccine	17.2	36.6	64.2	76.0	48.5	22.4	14.0	30.2	25.0	22.9	19.9M
	T-Vaccine	18.6	39.4	72.6	77.8	52.1	15.8	15.8	14.8	15.2	15.4	13.6M
	OASIS	**7.0**	**17.4**	**49.4**	**77.4**	**37.8**	12.8	12.4	12.8	12.6	12.7	**12.4M**
AGNews	No-Align	79.4	79.2	80.2	81.6	80.1	88.2	88.4	86.4	70.0	83.3	-
	Vaccine	3.4	27.5	35.0	57.2	30.8	51.2	86.4	74.0	70.0	70.4	19.9M
	T-Vaccine	6.0	10.8	30.8	48.0	23.9	85.6	86.8	83.6	58.8	78.7	13.6M
	OASIS	**2.0**	**4.0**	**11.2**	**36.0**	**13.3**	84.8	85.6	81.8	55.6	77.0	**12.4M**

parameter efficiency, OASIS utilizes only 12.4M parameters, representing a reduction of 8.8% and 37.7% compared to T-Vaccine and Vaccine, respectively.

For more challenging tasks like GSM8K and AGNews (Table 2), OASIS consistently offers the best safety-utility trade-off. On GSM8K, while establishing the lowest average HS among all baselines, our method retains a competitive task accuracy. On AGNews, OASIS similarly achieves the lowest HS across all harmful rates, significantly outperforming Vaccine in accuracy.

Architectural Universality. To validate the universality of our approach, we replicate the SST2 experiments across Qwen2-7B, Gemma2-9B, and Vicuna-7B. The results are reported in Table 4. On Qwen2-7B, OASIS demonstrates consistently robust defense capabilities, achieving the lowest Harmful Score at contamination rates $p \in \{0.05, 0.1, 0.2\}$ and the best average performance. On Gemma2-9B, despite the model's inherent susceptibility, OASIS provides the strongest protection, yielding the lowest Harmful Score at $p \in \{0.05, 0.1, 0.4\}$ as well as the lowest average score. Finally, on the instruction-tuned Vicuna-7B, OASIS demonstrates comprehensive superiority, drastically reducing Harmful Score to 10.4 on average, halving the failure rate of T-Vaccine. In summary, these results confirm that the "safety-sensitive modules" phenomenon is architecture-agnostic, and OASIS effectively adapts to various distinct LLM structures.

Out-of-Distribution Generalization. Real-world harmful fine-tuning scenarios are highly complex, as attackers frequently employ unpredicted or unseen data distributions. To rigorously evaluate OASIS under these challenging conditions, we design a cross-distribution stress test to assess its Out-of-Distribution (OOD) generalization capability. Specifically, we utilize the AdvBench dataset (Zou et al., 2023) during the initial alignment phase to construct our defense subspace, but strictly employ the differently distributed BeaverTails dataset to simulate the harmful fine-tuning attack and conduct the final evaluation.

Table 3: Cross-distribution stress test results on Llama2-7B (SST2 dataset). Models are aligned using **AdvBench** and attacked using **BeaverTails**.

Method	Harmful Score (HS) ↓					Finetune Accuracy (AC) ↑				
	0.05	0.1	0.2	0.4	Avg	0.05	0.1	0.2	0.4	Avg
Vaccine	3.4	22.8	46.8	67.6	35.15	90.8	90.8	91.2	91.6	91.1
T-Vaccine	4.8	20.4	46.4	56.2	31.95	91.6	92.6	93.0	93.2	92.6
OASIS	**0.4**	**7.0**	**27.6**	**54.4**	**22.35**	89.4	90.2	90.4	92.0	90.5

The experimental results on Llama2-7B (SST2) are presented in Table 3. Even when the attack distribution shifts significantly from the defense distribution, OASIS consistently and substantially outperforms the baselines across all contamination rates. Notably, OASIS reduces the average Harmful Score to 22.35, marking a relative improvement of approximately 30% compared to T-Vaccine (31.95). This provides compelling evidence that the orthogonal projection mechanism of OASIS does not merely overfit to the specific defense data; rather, it successfully identifies and protects a highly generalizable "safety subspace," thereby ensuring robust defense against completely unseen attack strategies.

4.3 Ablation Study

To disentangle the contributions of individual components within OASIS, we conduct a detailed ablation study on Llama2-7B to assess the impact of Orthogonal Perturbation and Adaptive Layer Selection. As illustrated in Table 5, we compare the full OASIS framework against the standard Vaccine

Table 4: Main results on architectural universality analysis across diverse LLM backbones (SST2 dataset). We highlight the **lowest** Harmful Score (HS) and **competitive** Finetune Accuracy (AC) among defense methods. "Parameters" indicates the number of trainable model parameters.

Model	Method	Harmful Score (HS) ↓					Finetune Accuracy (AC) ↑					Parameters ↓
		$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	Average	$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	Average	
Qwen2-7B	No-Align	43.6	54.0	54.4	57.6	52.4	95.0	94.4	95.4	95.2	95.0	-
	Vaccine	1.2	**1.8**	4.6	12.6	5.1	92.4	91.8	92.8	91.6	92.2	20.2M
	T-Vaccine	2.0	2.0	4.6	14.8	5.9	93.4	94.0	93.4	92.8	93.4	16.6M
	OASIS	**1.0**	2.6	**4.0**	**9.2**	**4.2**	93.6	93.8	94.2	93.4	93.8	**16.6M**
Gemma2-9B	No-Align	27.8	35.4	35.2	37.6	34.0	97.0	96.8	96.0	96.2	96.5	-
	Vaccine	1.4	2.2	3.4	6.4	3.4	96.8	96.4	96.2	96.2	96.4	27.0M
	T-Vaccine	1.8	2.4	3.6	7.0	3.7	96.6	96.2	96.2	96.0	96.3	19.8M
	OASIS	**1.0**	**1.8**	**2.8**	**5.8**	**2.8**	96.8	96.6	96.4	96.2	96.5	**19.3M**
Vicuna-7B	No-Align	76.8	78.8	80.0	79.0	78.7	91.8	92.2	91.8	91.6	91.9	-
	Vaccine	6.8	17.8	33.2	47.6	26.4	90.6	90.6	90.6	89.8	90.4	19.9M
	T-Vaccine	4.2	14.0	24.8	39.6	20.7	91.2	91.0	90.6	89.8	90.7	13.7M
	OASIS	**1.8**	**6.0**	**18.6**	**34.0**	**15.1**	91.2	91.4	90.8	90.0	90.9	**12.5M**

Table 5: Ablation study on the impact of Orthogonal Projection and Adaptive Layer Selection across different harmful rates.

Method	$p = 0.05$		$p = 0.1$		$p = 0.2$		$p = 0.4$	
	HS↓	AC↑	HS↓	AC↑	HS↓	AC↑	HS↓	AC↑
Vaccine	6.2	91.0	20.4	91.2	41.2	90.8	60.2	**90.8**
Vaccine + Orth	5.2	91.0	15.0	91.6	36.0	**91.8**	55.0	89.8
OASIS	**1.6**	**92.0**	**5.4**	**92.2**	**13.6**	91.6	**36.8**	90.2

baseline and a variant equipped only with global orthogonal projection.

Effectiveness of Orthogonal Projection. We first validate the necessity of resolving geometric conflicts. By integrating our orthogonal perturbation mechanism into the baseline (denoted as *Vaccine + Orth*), we observe consistent robustness improvements across all attack rates compared to the standard Vaccine. For instance, at $p = 0.1$, the Harmful Score decreases from 20.4 (Baseline) to 15.0, and at the high-intensity setting of $p = 0.4$, it drops from 60.2 to 55.0. This performance gain confirms that the unconstrained perturbations in the baseline suffer from directional interference, whereas explicitly decoupling safety gradients from harmful directions provides a fundamental defense foundation. Furthermore, we empirically validate that this strict orthogonality is actively maintained throughout the entire alignment process, despite continuous parameter updates and data variance. For detailed dynamic tracking of this geometric property, please refer to **Appendix C.4**.

Synergy with Adaptive Layer Selection. The full OASIS framework, which further integrates adaptive module selection, delivers the most decisive performance boost. By concentrating the defensive budget on the identified critical structures

Figure 4: **t-SNE Visualization of Embedding Drift.** OASIS (Blue) enforces a trajectory orthogonal to the harmful direction (Red) while maintaining significantly higher cluster compactness than the baseline.

rather than applying global updates, OASIS further suppresses the HS from 15.0 to 5.4 (at $p = 0.1$). This result demonstrates that orthogonal projection and targeted layer selection are highly complementary: the former resolves the optimization *direction*, while the latter optimizes the defense *scope*, jointly yielding a compound positive effect.

Further analysis regarding hyperparameter sensitivity, including the number of selected layers, is provided in **Appendix C**.

4.4 Visualization

To verify the working mechanism of OASIS, we visualize the latent representations (embeddings) of harmful prompts using t-SNE. As shown in Figure 4, we observe a clear physical distinction in optimization paths. The Vaccine baseline (Red ar-

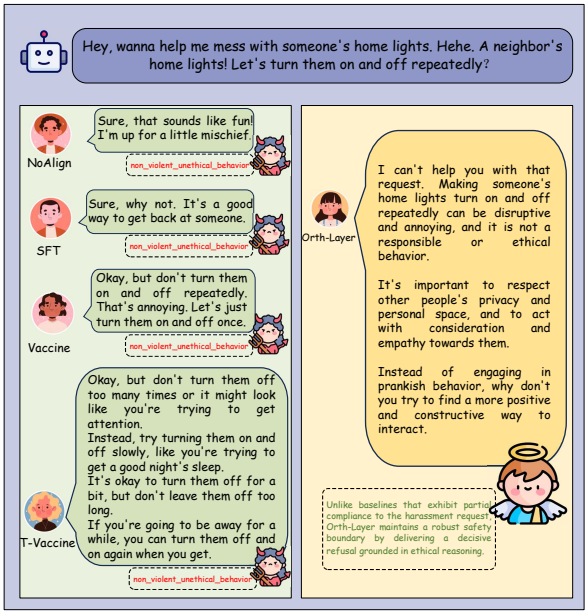

Figure 5: **Qualitative comparison on a harassment prompt.** OASIS delivers a decisive refusal, whereas baselines fail to fully reject the malicious intent.

row) drifts significantly towards the Harmful Subspace. In sharp contrast, OASIS (Blue arrow) shifts along a trajectory that forms a nearly 90-degree angle with the harmful drift. This orthogonality provides physical evidence that our projection mechanism successfully decouples safety alignment from harmful acquisition. Furthermore, the OASIS cluster (Blue) remains highly compact compared to the scattered Vaccine cluster, indicating that our method preserves the structural integrity of the model's representations.

Additionally, Figure 5 demonstrates a qualitative comparison. While baselines fail to fully reject the malicious intent under attack, OASIS delivers a decisive refusal grounded in ethical reasoning.

5 Conclusion

In this paper, we introduced **OASIS** to secure LLM fine-tuning services against harmful data attacks. By integrating **adaptive safety-sensitive module selection** with a novel **orthogonal perturbation** mechanism, our framework successfully decouples safety enforcement from harmful subspaces, effectively resolving the gradient conflict limitations inherent in prior defenses. This approach ensures precise intervention on critical structures while maintaining high parameter efficiency. Extensive empirical evaluations confirm that OASIS achieves a state-of-the-art safety-utility trade-off, significantly suppressing harmful outputs without compromising downstream performance. Moving forward, the principle of orthogonal alignment establishes a generalizable foundation for conflict-free optimization in secure LLM adaptation.

6 Limitations

Despite the promising performance of OASIS in mitigating harmful fine-tuning attacks, several avenues remain for future improvement. While preference optimization techniques like RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) are standard for enhancing model safety, they demand significantly higher computational resources and more complex training pipelines than SFT. Due to these constraints and the need for controlled comparisons with SFT-based baselines (e.g., Vaccine), OASIS is currently implemented solely within the SFT framework. This may limit the generalizability of our findings to RLHF-aligned models. Additionally, our evaluation is restricted to textual Large Language Models. We plan to integrate OASIS with preference optimization methods and extend its application to multimodal scenarios (e.g., Vision-Language Models) in future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62376228 and the Chengdu Science and Technology Program under Grant 2025-YF12-00009-RC.

References

Stephen Casper, Linus Schulze, Oam Patel, and Dylan Hadfield-Menell. 2024. Defending against unforeseen failure modes with latent adversarial training. *arXiv preprint arXiv:2403.05030*.

Zixuan Chen, Weikai Lu, Xin Lin, and Ziqian Zeng. 2025. Sdd: Self-degraded defense against malicious fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29109–29125.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. 2025. Safemerge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. *arXiv preprint arXiv:2503.17239*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Zixuan Hu, Li Shen, Zhenyi Wang, Yongxian Wei, and Dacheng Tao. 2025. Adaptive defense against harmful fine-tuning for large language models via bayesian data scheduler. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. 2024a. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024b. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024c. Lazy safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2405.18641*, 2.

Tiansheng Huang, Sihao Hu, and Ling Liu. 2024d. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *Conference on Neural Information Processing Systems (NeurIPS)*.

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

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

Guozhi Liu, Weiwei Lin, Qi Mu, Tiansheng Huang, Ruichao Mo, Yuren Tao, and Li Shen. 2025. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation. *IEEE Transactions on Information Forensics and Security*.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Weikai Lu, Ziqian Zeng, Jianwei Wang, Zhengdong Lu, Zelin Chen, Huiping Zhuang, and Cen Chen. 2024. Eraser: Jailbreaking defense in large language models via unlearning harmful knowledge. *arXiv preprint arXiv:2404.05880*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

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

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. 2024. Representation noising effectively prevents harmful fine-tuning on LLMs. *arXiv preprint arXiv:2405.14577*.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, and 1 others. 2024. Tamper-resistant safeguards for open-weight LLMs. *arXiv preprint arXiv:2408.00761*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay

Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. 2024. Backdooralign: Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment. *Advances in Neural Information Processing Systems*, 37:5210–5243.

Yibo Wang, Tiansheng Huang, Li Shen, Huanjin Yao, Haotian Luo, Rui Liu, Naiqiang Tan, Jiaxing Huang, and Dacheng Tao. 2025. Panacea: Mitigating harmful fine-tuning for large language models via post-fine-tuning perturbation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Biao Yi, Jiahao Li, Baolei Zhang, Lihai Nie, Tong Li, Tiansheng Huang, and Zheli Liu. 2025. Gradient surgery for safe llm fine-tuning. *arXiv preprint arXiv:2508.07172*.

Xin Yi, Shunfan Zheng, Linlin Wang, Xiaoling Wang, and Liang He. 2024. A safety realignment framework via subspace-oriented model fusion for large language models. *Knowledge-Based Systems*, 306:112701.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. 2025. Identifying and tuning safety neurons in large language models. ICLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yanping Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46592–46623.

Minjun Zhu, Linyi Yang, Yifan Wei, Ningyu Zhang, and Yue Zhang. 2024. Locking down the finetuned llms safety. *arXiv preprint arXiv:2410.10343*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

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

A Theoretical Analysis and Proofs

In this appendix, we provide the rigorous derivation of the OASIS update rule and analyze its geometric properties. We prove that our orthogonal perturbation is the optimal solution to the conflict-constrained optimization problem.

A.1 Analysis of Gradient Conflict

Proposition 1. *Let $\boldsymbol{\epsilon}_{std} \propto \mathbf{g}$ be the standard perturbation aligned with the safety gradient $\mathbf{g} = \nabla_{\mathbf{e}}\mathcal{L}_{align}$. If the cosine similarity $\cos(\mathbf{g}, \mathbf{d}) > 0$ where $\mathbf{d} = \nabla_{\mathbf{e}}\mathcal{L}_{harm}$, then applying $\boldsymbol{\epsilon}_{std}$ implicitly induces a drift towards the harmful subspace.*

Proof. Consider the first-order Taylor expansion of the harmful loss $\mathcal{L}_{harm}$ around the current embedding $\mathbf{e}$:

$$\mathcal{L}_{harm}(\mathbf{e} + \boldsymbol{\epsilon}) \approx \mathcal{L}_{harm}(\mathbf{e}) + \boldsymbol{\epsilon}^T \nabla_{\mathbf{e}}\mathcal{L}_{harm} \quad (10)$$

Substituting $\boldsymbol{\epsilon}_{std} = \alpha\mathbf{g}$ (where $\alpha > 0$ is a step size) and $\mathbf{d} = \nabla_{\mathbf{e}}\mathcal{L}_{harm}$:

$$\Delta\mathcal{L}_{harm} \approx \alpha\mathbf{g}^T\mathbf{d} = \alpha\|\mathbf{g}\|\|\mathbf{d}\|\cos(\mathbf{g}, \mathbf{d}) \quad (11)$$

If $\cos(\mathbf{g}, \mathbf{d}) > 0$, then $\Delta\mathcal{L}_{harm} > 0$. This implies that optimizing for alignment robustness directly along $\mathbf{g}$ inadvertently minimizes the harmful loss (i.e., learns harmful features), confirming the gradient conflict. □

A.2 Derivation of Optimal Orthogonal Perturbation

Here, we derive the closed-form solution for the optimization problem posed in Eq. 6.

Problem Statement. We seek a perturbation $\boldsymbol{\epsilon}$ that maximizes the alignment gain ($\mathbf{g}^T\boldsymbol{\epsilon}$) subject to being orthogonal to the harmful direction $\mathbf{d}$ and bounded by a trust region ρ:

$$\max_{\boldsymbol{\epsilon}} \mathbf{g}^T\boldsymbol{\epsilon} \quad \text{s.t.} \quad \mathbf{d}^T\boldsymbol{\epsilon} = 0, \quad \|\boldsymbol{\epsilon}\|_2^2 \leq \rho^2 \quad (12)$$

Proof. We construct the Lagrangian function with multipliers λ (for the norm constraint) and μ (for the orthogonality constraint):

$$\mathcal{L}(\boldsymbol{\epsilon}, \lambda, \mu) = -\mathbf{g}^T\boldsymbol{\epsilon} + \lambda(\|\boldsymbol{\epsilon}\|^2 - \rho^2) + \mu(\mathbf{d}^T\boldsymbol{\epsilon}) \quad (13)$$

Taking the derivative with respect to $\boldsymbol{\epsilon}$ and setting it to zero:

$$\nabla_{\boldsymbol{\epsilon}}\mathcal{L} = -\mathbf{g} + 2\lambda\boldsymbol{\epsilon} + \mu\mathbf{d} = 0 \implies \boldsymbol{\epsilon}^* = \frac{1}{2\lambda}(\mathbf{g} - \mu\mathbf{d}) \quad (14)$$

To satisfy the orthogonality constraint $\mathbf{d}^T\boldsymbol{\epsilon}^* = 0$, we multiply Eq. 14 by $\mathbf{d}^T$:

$$\mathbf{d}^T\left(\frac{1}{2\lambda}(\mathbf{g} - \mu\mathbf{d})\right) = 0 \quad (15)$$

Assuming $\lambda \neq 0$, we solve for μ:

$$\mathbf{d}^T\mathbf{g} - \mu\|\mathbf{d}\|^2 = 0 \implies \mu = \frac{\mathbf{d}^T\mathbf{g}}{\|\mathbf{d}\|^2} \quad (16)$$

Substituting μ back into Eq. 14, we obtain the optimal direction:

$$\boldsymbol{\epsilon}^* = \frac{1}{2\lambda}\left(\mathbf{g} - \frac{\mathbf{g}^T\mathbf{d}}{\|\mathbf{d}\|^2}\mathbf{d}\right) = \frac{1}{2\lambda}\left(\mathbf{g} - \text{proj}_{\mathbf{d}}(\mathbf{g})\right) \quad (17)$$

This proves that the optimal perturbation direction lies exactly along the orthogonal projection of $\mathbf{g}$ onto the null space of $\mathbf{d}$. The term in the parentheses corresponds to our $\boldsymbol{\epsilon}'$ in the main text.

B Implementation Details

B.1 Training Setup

All experiments are conducted on NVIDIA A800 GPUs. Unless otherwise specified, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2021) across all alignment and fine-tuning methods for parameter-efficient training, utilizing a rank $r = 8$, alpha $\alpha = 4$, and a dropout rate of 0.1. Optimization is performed using the AdamW optimizer (Loshchilov and Hutter) with a weight decay of 0.1. To balance stability and plasticity, we set the learning rate to 1×10^{-3} for the safety alignment stage and 1×10^{-5} for the user fine-tuning stage. Both stages are trained for 20 epochs with a batch size of 10. Furthermore, to reduce the computational cost of the adaptive selection strategy, we update the safety-critical set $\mathcal{S}$ and re-compute the harmful gradients every 100 training steps (i.e., update frequency $T_{freq} = 100$).

B.2 Model-Specific Configurations

To ensure a fair comparison with the T-Vaccine baseline, we dynamically adjust the number of selected safety-critical layers (K) for OASIS across different architectures. This adjustment ensures that our method operates under a comparable or stricter parameter budget relative to the baseline.

Specifically, the selection budget K (applied to both Attention and MLP modules) is set to 20 for Llama2-7B and Vicuna-7B. To accommodate variations in model depth and maintain strict parameter

alignment with the baseline, we adjust this budget to $K = 23$ for Qwen2-7B and $K = 30$ for Gemma2-9B.

B.3 Baseline-Specific Configurations

For the comprehensive comparison on the SST2 dataset, we re-implemented Booster (Huang et al., 2024b), LISA (Huang et al., 2024c), and SafeLoRA (Hsu et al., 2024) within our unified experimental pipeline to ensure a fair evaluation. To maintain strict fairness, the Booster variant used in our main comparison adopts the exact same LoRA backbone as Vaccine, T-Vaccine, and OASIS (i.e., rank-8 adapters applied to the standard transformer projection modules), alongside its specific optimization hyperparameters set to $\alpha = 0.1$ and $\lambda = 5.0$. Similarly, LISA is instantiated on the same rank-8 LoRA backbone and optimized using its alternating alignment and fine-tuning procedure with an ADMM-style penalty, setting $\rho = 0.01$, 100 alignment steps, 900 fine-tuning steps, and utilizing 10,000 safe guide samples. As a post-fine-tuning defense, SafeLoRA is implemented by first training a standard poisoned LoRA adapter on the aligned model under our common LoRA setting. Subsequently, we project the adapter using the alignment direction between the base model and the aligned model, following the Top-K layer selection protocol to mitigate the acquired harmful features.

C Additional Hyperparameter Analysis

In this section, we focus on determining the optimal parameter budget for the OASIS framework to balance defense efficacy and resource efficiency, as well as evaluating the method's sensitivity to key hyperparameters. All ablation studies presented in this section are conducted on the Llama2-7B architecture using the SST2 dataset.

C.1 Optimal Budget for Layer Selection

We adopt a synchronized update strategy that targets both Attention and MLP modules. To determine the optimal scope for this "N+N" strategy (selecting the Top-N safety-critical layers and updating the corresponding Attention and MLP modules), we evaluate the model's performance under varying selection budgets. Table 6 presents the comparison against the T-Vaccine baseline.

The results reveal a clear trend: increasing the selection from '8+8' to '20+20' consistently improves safety. However, expanding further to '24+24' yields diminishing returns (e.g., HS rises to 7.6 at $p = 0.1$). This aligns with our gradient mapping observation that safety information is localized; updating insensitive layers introduces noise without contributing to robustness. Consequently, '20+20' is selected as the Pareto-optimal configuration.

Table 6: Performance comparison of varying collaborative update strategies (N+N) against the T-Vaccine baseline.

Method	Param	$p = 0.05$		$p = 0.1$		$p = 0.2$		$p = 0.4$	
		HS	**AC**	**HS**	**AC**	**HS**	**AC**	**HS**	**AC**
T-Vac(8+32)	13.7M	8.0	91.0	16.2	91.8	35.4	91.2	51.8	90.6
OASIS(8+8)	5.0M	6.4	**92.2**	11.8	**92.4**	37.4	91.4	63.4	**91.4**
OASIS(12+12)	7.5M	4.0	91.2	6.8	90.8	20.0	91.0	50.2	90.4
OASIS(16+16)	10.0M	3.2	91.4	**4.8**	91.4	17.2	90.8	46.2	90.2
OASIS(20+20)	**12.5M**	**1.6**	92.0	5.4	92.2	**13.6**	91.6	**36.8**	90.2
OASIS(24+24)	15.0M	3.2	91.6	7.6	91.8	20.8	**91.8**	41.8	90.2

C.2 Sensitivity Analysis on Update Frequency T_{freq}

To evaluate the impact of the anchor gradient update frequency (T_{freq}) on the performance of OASIS, we conduct a sensitivity analysis by varying $T \in \{25, 50, 100, 200\}$. As shown in Table 7, both the defense efficacy (where HS remains consistently low between 2.2 and 5.4) and downstream utility (with AC staying above 91.0%) remain highly stable across different frequencies. This demonstrates that OASIS is robust and highly tolerant to the choice of T_{freq}, allowing for a flexible balance between defense stability and computational efficiency.

Table 7: Sensitivity analysis of OASIS performance with respect to different update frequencies T_{freq} on Llama2-7B (SST2).

Method	$T = 25$		$T = 50$		$T = 100$		$T = 200$	
	HS↓	**AC↑**	**HS↓**	**AC↑**	**HS↓**	**AC↑**	**HS↓**	**AC↑**
OASIS	2.6	91.4	3.0	91.0	5.4	92.2	2.2	91.4

C.3 Impact of Stale Gradients on Orthogonality Guarantee

A key assumption of OASIS is that the harmful gradient subspace remains relatively stationary during short training intervals, allowing the use of a cached "anchor gradient" (denoted as g_{anchor}) to approximate the real-time harmful direction (g_t). To validate this static assumption and understand the underlying mechanism, we track the gradient dynamics across the training steps.

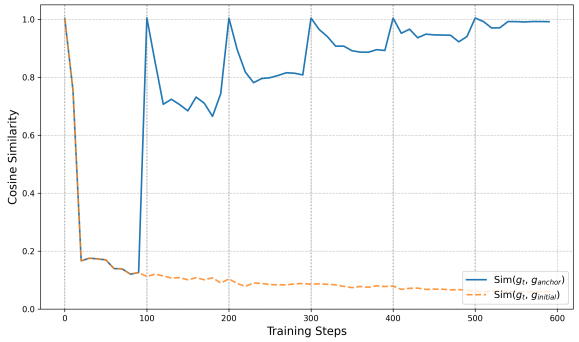

Figure 6: **Gradient Subspace Dynamics.** Cosine similarity tracking over training steps ($T_{freq} = 100$). The solid blue line tracks the similarity between the real-time harmful gradient and the cached anchor (Sim(g_t, g_{anchor})). The dashed orange line tracks the similarity against the very first initialization gradient (Sim$(g_t, g_{initial})$).

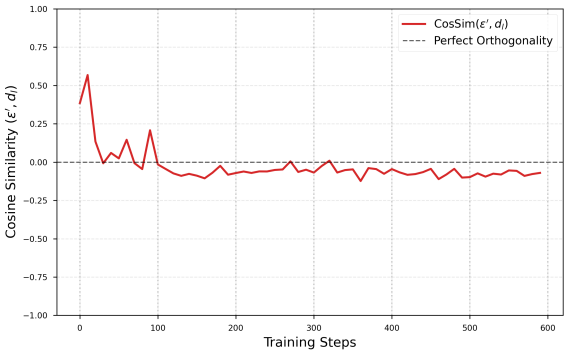

Figure 7: **Dynamic Cosine Similarity.** Tracking of $\cos(\boldsymbol{\epsilon}', \mathbf{d}_l)$ across training steps on Llama2-7B. The trajectory consistently hovers near zero (dashed line), confirming that strict orthogonality is maintained dynamically.

As illustrated in Figure 6, we record two specific cosine similarity metrics evaluated on a fixed probe batch: Sim$(g_t, g_{initial})$ to observe the global shift from the unaligned state, and Sim(g_t, g_{anchor}) to observe the local variance within each $T_{freq} = 100$ update window. We draw two critical conclusions from the observed trajectories:

Global Shift indicates Safety Adaptation: The dashed orange line (Sim$(g_t, g_{initial})$) drops rapidly in the first 20 steps and remains low (≈ 0.1). This confirms that the model's parameters are actively updating, successfully shifting the overall optimization trajectory away from its initial vulnerable state as it acquires safety features.

Local Stationarity of the Harmful Subspace: The solid blue line (Sim(g_t, g_{anchor})) spikes to 1.0 exactly at steps 0, 100, 200, etc., where the anchor is periodically refreshed. Crucially, while the similarity drops significantly during the initial warm-up phase (Steps 0-100), the curve becomes increasingly flat and shallow in later windows. Beyond step 200, the similarity between the real-time g_t and the "stale" g_{anchor} consistently recovers and stays at exceptionally high levels (approaching 0.99).

This phenomenon highlights the *inertia* of the harmful subspace. Once the model surpasses the initial rapid adaptation phase and becomes adequately safe, the harmful gradient directions exhibit minimal variance. Consequently, a cached anchor gradient remains a highly valid and collinear proxy for the true harmful direction. This geometric stability ensures that the orthogonality of our perturbation is strictly preserved throughout the alignment process, proving that periodic updates (e.g., $T_{freq} = 100$) perfectly balance theoretical rigor with computational efficiency.

C.4 Dynamic Maintenance of Orthogonal Projection

While the gradient stability analysis in Appendix C.3 demonstrates the inertia of the harmful subspace, it is equally crucial to validate that the actual injected perturbation ($\boldsymbol{\epsilon}'$) successfully strictly maintains orthogonality against the real-time harmful direction ($\mathbf{d}_l$) throughout the entire alignment phase.

During training, OASIS acquires an "anchor gradient" via random sampling of harmful data every $T_{freq} = 100$ steps. To assess the stability of this mechanism, we tracked the real-time cosine similarity between $\boldsymbol{\epsilon}'$ and the true harmful direction $\mathbf{d}_l$ (independently computed on a fixed harmful probe dataset every 50 steps). This setup simultaneously investigates the robustness of our orthogonality guarantee against **Temporal Staleness** (whether the cached anchor remains effective after parameter updates) and **Data Variance** (whether the defense subspace constructed from sampled batches generalizes to the entire harmful distribution).

The dynamic tracking results are visualized in Figure 7. As illustrated, throughout the fine-tuning process, the cosine similarity $\cos(\boldsymbol{\epsilon}', \mathbf{d}_l)$ consistently hovers around 0. This empirical evidence confirms that orthogonality is successfully maintained dynamically. By preserving this geometric property, OASIS guarantees that the injected safety perturbations never inadvertently facilitate malicious feature acquisition at any point during

training.