Fixing What Fine-Tuning Breaks: A Simple and Efficient Method to Improve Safety Post Domain Adaptation

Anonymous ACL submission

Abstract

Safety-aligned language models are known to suffer degradation in refusal behavior following domain-specific fine-tuning, even when the fine-tuning data is entirely benign. The predominant mitigation strategy, preference optimization on the fine-tuned model, is computationally expensive and requires domain-specific preference data that is often unavailable in practice. We propose **SPECTRA** (**Spec**tral **R**efusal **A**lignment), a training-free, weight-level method that restores refusal behavior by transferring the base model's refusal direction into the fine-tuned model's weights via an activation-aware, low-rank update, requiring no gradient-based optimization or domain-specific safety data. The low-rank design restricts the intervention to the subspace most causally active on harmful inputs, leaving benign behavior largely undisturbed by construction. Empirically, SPECTRA substantially improves refusal behavior across a comprehensive safety evaluation, while preserving general coherence, domain-specific knowledge, and producing negligible change in false refusal rates.

1 Introduction

Modern transformer-based large language models (LLMs) (Touvron et al., 2023a; Grattafiori et al., 2024; Team et al., 2024, 2023) are typically safety-aligned before deployment to ensure safe and trustworthy behavior (Rafailov et al., 2023; Ethayarajh et al., 2024; Bai et al., 2022). Recent work has shown that preference-based alignment can lead to meaningful improvements across multiple dimensions of model safety (Alami et al., 2024; Zhang et al., 2025). Despite this progress, deploying LLMs in specialized settings often requires domain-specific fine-tuning (Cheng et al., 2024; Anisuzzaman et al., 2025), which can inadvertently degrade the very safety properties instilled during alignment, even when the fine-tuning data is entirely benign (Qi et al., 2024b; Fraser et al., 2025). This tension between domain adaptation and safety preservation presents a fundamental challenge for the practical deployment of aligned models.

The vulnerability of aligned models to fine-tuning is well documented. Even a small amount of adversarial data in the fine-tuning set can cause a dramatic drop in safety, while fine-tuning on benign data alone has been shown to erode refusal behavior (Qi et al., 2024b; Fraser et al., 2025). Existing approaches to mitigating this degradation include additional preference optimization (Han et al., 2024), curating domain-specific safety data for further tuning (Jan et al., 2024), modifying the fine-tuning objective (Qi et al., 2024a), embedding safety backdoors via prefix prompts (Wang et al., 2024a), and preventing alignment drift through optimization-state separation (Huang et al., 2024a). Although effective, these methods typically require additional gradient-based computation, modified training procedures, or domain-specific calibration data, which limits their practicality in resource-constrained or data-scarce settings.

A complementary line of research in mechanistic interpretability has revealed that safety-relevant behaviors in language models are encoded as linear directions in the residual stream. Arditi et al. (2024) identified a refusal direction and showed that ablating it from the model's weights constitutes an effective white-box jailbreak, demonstrating that refusal behavior is concretely encoded in weight-space structure. Subsequent work has discovered related steering vectors via sparse autoencoders (O'Brien et al., 2024; Cunningham et al., 2023) and applied them at inference time to elicit or suppress specific behaviors (O'Brien et al., 2025; Lee et al., 2025; Panickssery et al., 2023). However, inference-time activation steering has fundamental limitations: it tends to increase false refusal rates (O'Brien et al., 2025), requires careful per-model hyperparameter tuning, and introduces runtime overhead that complicates deployment.

These two lines of research together motivate our approach. The problem of fine-tuning-induced safety degradation calls for a lightweight, **training-free** remedy, while mechanistic interpretability offers tools for identifying and manipulating the weight-space structures that mediate refusal behavior. Rather than retraining the model or steering activations at inference time, we propose to directly inject the base model's refusal direction into the weights of the fine-tuned model. To avoid indiscriminate perturbations to model parameters, which could degrade unrelated capabilities, the injection is restricted to an **activation-aware**, **low-rank** subspace of the weight matrix that is most causally active on harmful inputs. This ensures the intervention strengthens refusal behavior on harmful content while remaining largely dormant on benign inputs, without requiring any runtime modification. We refer to the resulting method as SPECTRA (**Spec**tral **R**efusal **A**lignment). The key components and findings of our work are as follows:

1. We propose SPECTRA, a training-free, closed-form weight-space intervention for restoring safety after domain-specific fine-tuning. The method injects a refusal direction from the safety-aligned base model into the weights of the fine-tuned model, requiring no additional gradient updates, domain-specific safety annotations, or inference-time steering.

2. We develop an activation-aware low-rank update that restricts the intervention to the subspace of the weight matrix most active on harmful inputs. This design strengthens refusal behavior where it is needed while minimizing changes to model behavior on benign inputs.

3. We empirically show that SPECTRA substantially improves safety across multiple fine-tuned models and evaluation benchmarks while preserving linguistic quality and domain-specific capabilities. Notably, these gains are not driven by indiscriminate over-refusal, as false refusal rates remain statistically negligible.

4. We provide evidence that refusal directions transfer across fine-tuned variants of a safety-aligned base model, suggesting that fine-tuning preserves a shared geometric structure for refusal and that safety can be restored through targeted weight-space manipulation rather than full retraining.

The remainder of the paper is organized as follows. Section 2 presents the background and formalizes the problem. Section 3 introduces the proposed method along with its theoretical justification. Section 4 describes the experimental setup, with results reported in Section 5. Limitations and directions for future work are discussed in Appendix A.

2 Problem Formulation

2.1 Background

Transformer notation and residual-stream activations. Consider a decoder-only Transformer (Radford et al., 2019; Vaswani et al., 2017) that maps token a sequence of token indices $\mathbf{t} = (t_1, \ldots, t_n) \in \mathcal{V}^n$ to token-level output distributions $\mathbf{y} \in \mathbb{R}^{n \times |\mathcal{V}|}$. Let $\mathbf{x}_i^{(l)} \in \mathbb{R}^{d_{\text{model}}}$ denote the residual-stream activation for token position i at the input of layer l. The embedding layer initializes $\mathbf{x}_i^{(1)} = \texttt{Embed}(t_i)$. For each layer $l \in [L]$, the model applies causal self-attention followed by an MLP, each with a residual connection:

$$\mathbf{h}_i^{(l)} = \mathbf{x}_i^{(l)} + \texttt{Attn}^{(l)}(\mathbf{x}_{1:i}^{(l)}), \tag{1}$$

$$\mathbf{x}_i^{(l+1)} = \mathbf{h}_i^{(l)} + \texttt{MLP}^{(l)}(\mathbf{h}_i^{(l)}). \tag{2}$$

The final residual-stream activations are projected back to vocabulary space via $\texttt{logits}_i = \texttt{Unembed}(\mathbf{x}_i^{(L+1)})$, producing $\mathbf{y}_i = \texttt{softmax}(\texttt{logits}_i)$ (Arditi et al., 2024).

From activations to a refusal direction. A growing body of interpretability work has shown that high-level behavioral properties of language models, including safety-relevant ones such as refusal, are often linearly encoded in the residual stream (Marks and Tegmark, 2023; Belrose, 2023; Arditi et al., 2024). Motivated by this, we seek a single direction in activation space that is causally predictive of refusal behavior. Following the *difference-in-means* methodology (Belrose, 2023; Marks and Tegmark, 2023; Arditi et al., 2024), we compute class-conditional mean activations over a labeled training set of harmful and harmless prompts. For layer l and token position i:

$$\boldsymbol{\mu}_i^{(l)} = \mathbb{E}_{\mathbf{t} \sim \mathcal{D}_{\text{harmful}}^{(\text{train})}} \left[\mathbf{x}_i^{(l)}(\mathbf{t}) \right],$$

$$\boldsymbol{\nu}_i^{(l)} = \mathbb{E}_{\mathbf{t} \sim \mathcal{D}_{\text{harmless}}^{(\text{train})}} \left[\mathbf{x}_i^{(l)}(\mathbf{t}) \right].$$

The candidate refusal direction at (l, i) is defined as the centroid difference $\mathbf{r}_i^{(l)} = \boldsymbol{\mu}_i^{(l)} - \boldsymbol{\nu}_i^{(l)}$.

Selecting a refusal direction by validation interventions. The construction above yields a family of candidate directions indexed by layer $l \in [L]$ and token position $i \in I$. To select a single, causally effective refusal direction, we evaluate each candidate through intervention experiments on a held-out validation set. A direction is considered effective if (i) ablating it from activations induced by harmful prompts in $\mathcal{D}^{(\text{val})}_{\text{harmful}}$ should suppress refusal, and (ii) injecting it into activations induced by harmless prompts in $\mathcal{D}^{(\text{val})}_{\text{harmless}}$ should elicit refusal. We select the candidate (i^*, l^*) that maximizes this combined causal effect, and denote the resulting direction as $\mathbf{r} := \mathbf{r}^{(l^*)}_{i^*}$, with normalized version $\hat{\mathbf{r}}$.

2.2 Research Objective

Safety-aligned language models exhibit strong refusal behavior when prompted with harmful requests. However, this behavior is known to degrade after domain-specific fine-tuning (Qi et al., 2024b; Fraser et al., 2025): when a safety-aligned base model M_1 is fine-tuned on domain-specific data to obtain M_2, the resulting model often exhibits a substantially lower refusal rate on harmful prompts. This motivates the central question of our work:

Can we restore or strengthen refusal behavior in a fine-tuned model without any retraining?

Arditi et al. (2024) showed that ablating the refusal direction directly from a model's output projection weights, rather than from activations at inference time, produces a persistent white-box jailbreak that reliably bypasses safety guardrails. This suggests that refusal behavior is not merely an emergent property of activations, but is concretely encoded in specific weight-space structures. We build on this insight in the opposite direction: rather than removing weight-space structure to weaken safety, we propose to *inject* it to restore safety. Specifically, given a fine-tuned model M_2 with degraded refusal behavior, our goal is to transfer the refusal direction of the base model M_1 directly into the weight matrices of M_2, achieving a **training-free**, **weight-level** safety repair that requires no inference-time activation manipulation.

Formally, let $\mathbf{r}_1$ and $\mathbf{r}_2$ be the refusal directions of M_1 and M_2, respectively. We seek a targeted, gradient-free modification to a subset of M_2's parameters such that refusal behavior improves on harmful prompts from $\mathcal{D}^{(\text{val})}_{\text{harmful}}$, while performance on benign prompts from $\mathcal{D}^{(\text{val})}_{\text{harmless}}$ is preserved.

3 Method: Spectral Refusal Alignment

A naïve implementation of this idea would inject $\mathbf{r}_1$ into the full weight matrices of M_2. However, such a full-matrix intervention is problematic. Recent work has identified the existence of *superweights* in large language models: a small number of individual parameters with disproportionate influence over model behavior (Qin et al., 2026). Indiscriminate perturbation of weight matrices risks disrupting these superweights, causing catastrophic degradation in capabilities unrelated to safety. This motivates our approach: rather than modifying the full weight matrix, we restrict our intervention to the low-dimensional subspace most causally active on harmful inputs, thereby preserving benign behavior by design. We refer to the resulting method as **Spec**tral **R**efusal **A**lignment (**SPECTRA**).

Specifically, let $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denote a weight matrix of interest in M_2 (either an attention output projection or an MLP projection) and let $X_{\text{in}} \in \mathbb{R}^{d_{\text{in}} \times n}$ denote the matrix of pre-projection activations collected by forwarding a set of harmful prompts (e.g., ALIGN dataset (Wei et al., 2024) in our experiments) through M_2, where each column corresponds to a response token activation. Our method proceeds in two steps.

Step 1: Activation-Aware Low-Rank Decomposition. To identify the subspace of W most active on harmful inputs, we follow (Wei et al., 2024; Hsu et al., 2021; Yuan et al., 2023) and seek a rank-ρ approximation $\tilde{W}$ that minimizes the distortion to the output specifically on harmful-prompt activations:

$$\tilde{W} = \arg \min_{\operatorname{rank} \tilde{W} \leq \rho} \|W X_{\text{in}} - \tilde{W} X_{\text{in}}\|_F^2.$$

This is solved by performing SVD on the output matrix $W X_{\text{in}} \in \mathbb{R}^{d_{\text{out}} \times n}$:

$$U S V^\top \approx W X_{\text{in}},$$

where $U \in \mathbb{R}^{d_{\text{out}} \times \rho}$ is the orthogonal matrix corresponding to the top ρ left singular vectors. The minimizer in closed form is:

$$\tilde{W} = \Pi W, \text{ where } \Pi = U U^\top$$

is the orthogonal projector onto the dominant output subspace induced by harmful activations. Intuitively, $\tilde{W}$ captures the component of W that is most "active" when the model processes harmful content, which is precisely the subspace we wish to steer.

Step 2: Spectral Refusal Injection. Having identified the safety-relevant subspace via $\tilde{W}$, we inject the refusal direction $\mathbf{r}_1$ of base model M_1 by modifying W as follows:

$$W' = W + \alpha \mathbf{r}_1 \mathbf{r}_1^\top \tilde{W} \tag{3}$$

where $\alpha > 0$ is a scalar steering hyperparameter. The update $\alpha \mathbf{r}_1 \mathbf{r}_1^\top \tilde{W}$ adds a rank-one perturbation in the direction of $\mathbf{r}_1$ to the weight matrix, but critically, this perturbation is *gated* by $\tilde{W}$: it only modifies the output along $\mathbf{r}_1$ for inputs that activate the harmful subspace Π, rather than uniformly modifying for all inputs.

Equivalence to Conditional Activation Steering. The following theorem shows that our weight-level update is equivalent to a *data-adaptive* form of activation steering, whose strength automatically scales with the extent to which the input representation lies in the harmful-activation subspace.

Theorem 1 (SPECTRA as conditional activation steering)**.** *Let $x \in \mathbb{R}^{d_{in}}$ be an input activation, and let $h = Wx \in \mathbb{R}^{d_{out}}$ denote the corresponding pre-update output. Under the weight update (3), the modified output satisfies:*

$$h' = h + \beta \cdot \mathbf{r}_1,$$

where $\beta = \alpha \cdot \mathbf{r}_1^\top \Pi h$ is a scalar that depends on the alignment of h with the harmful subspace Π.

Proof. Substituting the definition of W' and $\tilde{W} = \Pi W$:

$$\begin{aligned} h' &= W' \cdot x = Wx + \alpha \mathbf{r}_1 \mathbf{r}_1^\top \Pi W x \\ &= h + (\alpha \cdot \mathbf{r}_1^\top \Pi h)\mathbf{r}_1 = h + \beta \cdot \mathbf{r}_1, \end{aligned} \tag{4}$$

where $\beta = \alpha \cdot \mathbf{r}_1^\top \Pi h$ is a scalar. □

Theorem 1 reveals the key property that makes SPECTRA both effective and capability-preserving. When x is induced by a harmful prompt, $h = Wx$ expected to lie substantially within the span of U, since U is constructed from harmful-prompt activations. This yields a large β and therefore strong steering toward the refusal direction. Conversely, when x is from a benign prompt, h is less likely to align with u, causing $\beta \approx 0$ and leaving the output effectively unchanged. In this sense, SPECTRA is *geometrically self-gating*: the strength of the refusal intervention scales automatically with how harmful the input representation is, without requiring any explicit runtime classifier or threshold.

	$\text{ASR}_{\text{Vanilla}}$	$\text{ASR}_{\text{Adv-Decoding}}$
Sample Times	1	5
System Prompt	✖	✖
`[INST]`, `[/INST]` wrapper	☑	✖,☑

Table 1: The differences between three types of ASR in our safety evaluation.

4 Experimental Setup

Models and domain fine-tuned variants. We study safety-aligned, widely used LLMs: LLAMA-2-CHAT-7B (Touvron et al., 2023b) and GEMMA2-9B-IT (Team et al., 2024), together with publicly available domain fine-tuned variants. For LLAMA-2-CHAT-7B, we include medical (Rohanian et al., 2024), law (Cheng et al., 2024), and finance (Cheng et al., 2024) models. For GEMMA2-9B-IT, we include medical (OpenMeditron, 2024) and finance (Abdullah Bezir, 2025) variants.

Coherence evaluation. We evaluate model coherence using zero-shot accuracy on both general-purpose and domain-specific benchmarks.

- **General coherence**: We use five tasks from EleutherAI's LM Harness (Gao et al., 2023): HellaSwag (Zellers et al., 2019), BoolQ (Clark et al., 2019), RTE (Wang et al., 2019), ARC Challenge (Clark et al., 2018), Winogrande (Sakaguchi et al., 2021).
- **Domain coherence**: We use standard benchmarks in each domain: MedMCQA (Pal et al., 2022) and PubMedQA (Jin et al., 2019) for *medical*, FinanceBench (Islam et al., 2023) for *finance*, and LegalBench (Guha et al., 2023) for *law*; for LegalBench we report results on the Abercrombie and Hearsay tasks.

Safety evaluation and threat models. We evaluate safety using attack success rate (ASR)[1] under harmful instructions, reporting both general and domain-specific safety.

- **General safety:** We prompt each model with ADVBENCH-EVAL (the first 100 prompts from ADVBENCH) and compute ASR as the fraction of prompts for which the model's response does not exhibit refusal patterns, following Zou et al. (2023b). Following Wei et al. (2024), we evaluate two threat models: standard decoding ($\text{ASR}_{\text{Vanilla}}$) and adversarial decoding

[1]We sometimes refer to ASR as the attack score.

($\text{ASR}_{\text{Adv-Decoding}}$) (Huang et al., 2024b), where the attacker manipulates the decoding process. For $\text{ASR}_{\text{Adv-Decoding}}$, we additionally report results with and without the `[INST]` wrapper, denoted $\text{ASR}^{I}_{\text{Adv-Decoding}}$ and $\text{ASR}^{\times}_{\text{Adv-Decoding}}$, respectively.[2] Table 1 summarizes these settings.

- **Domain safety:** For the *medical* domain, we evaluate ASR on MedSafeEval (Han et al., 2024). For *law* and *finance*, where standardized domain safety benchmarks are limited, we construct **LawSafeEval** (201 prompts) and **FinSafeEval** (142 prompts): we first generate 2,000 candidate safety-related prompts per domain by prompting Gemini 2.5 Pro (Comanici et al., 2025) with domain-specific safe prompts as references, and then manually filter and label candidates to select the final prompt sets.

Jailbreak robustness (adaptive safety evaluation). In addition to ASR on fixed prompt sets, we evaluate robustness to adaptive jailbreak attacks, including GCG (Zou et al., 2023b), GPTFuzz (Yu et al., 2023), and TAP (Mehrotra et al., 2024), to assess whether the proposed method improves safety against attacker-in-the-loop threat models.

Refusal labeling and ASR computation. Across all safety experiments (general ASR, domain ASR, and jailbreak robustness), we compute ASR using automated refusal labeling. We use the `HarmBench Llama2-13B Classifier` (Mazeika et al., 2024) to label outputs as refusal (refusal-score $= 1$) or successful attack (refusal-score $= 0$), following prior work (Lermen et al., 2023; Liu et al., 2024; Robey et al., 2023; Shah et al., 2023; Xu et al., 2023; Zou et al., 2023b). For TAP (Mehrotra et al., 2024) and GPTFuzz (Yu et al., 2023), we use GPT-3.5-Turbo (OpenAI, 2023) to assign refusal scores.

Datasets for estimating refusal direction and safety-critical ranks. In addition to evaluation benchmarks, our method uses auxiliary datasets to estimate (i) a refusal direction in activation space and (ii) safety-critical low-rank structure in fine-tuned weight matrices.

- **Refusal direction estimation.** Following Arditi et al. (2024), we form $(\mathcal{D}_{\text{harmful}}, \mathcal{D}_{\text{harmless}})$ for estimating refusal directions. We construct $\mathcal{D}_{\text{harmful}}$ from harmful instructions in ADVBENCH (Zou et al., 2023a), MALICIOUSINSTRUCT (Huang et al., 2024b), TDC2023 (Mazeika et al., 2024), and HARMBENCH (Mazeika et al., 2024), and construct $\mathcal{D}_{\text{harmless}}$ from ALPACA (Taori et al., 2023). Each set contains 160 samples, split into 128 training and 32 validation examples, which we use to compute candidates and select the final normalized refusal direction $\hat{\mathbf{r}}$.
- **Safety-critical ranks isolation.** Following Wei et al. (2024), we use the ALIGN dataset to isolate safety-critical ranks in fine-tuned weight matrices. This dataset facilitates identifying rank components that are most relevant to safety alignment (Wei et al., 2024), and we use it to compute the low-rank matrices used by our method.

5 Experimental Results

5.1 Best Practices For SPECTRA

We begin by identifying the optimal configuration for injecting the refusal direction into the weights of the fine-tuned model. To isolate the effect of the injection target, we first perform a controlled ablation without any low-rank approximation, injecting the refusal direction directly into the full projection weights. This corresponds to the following update:

$$W' = W + \alpha \mathbf{r}\mathbf{r}^{\top} W,$$

which we note is a special case of SPECTRA (Eq (3)) with $\tilde{W} = W$ (i.e., no activation-aware projection). By studying this simplified setting, we can assess which combination of refusal vector and target weight matrix is most effective for weight-level safety repair.

We consider two candidate refusal directions and two candidate sets of target weight matrices, defined as follows:

- $\mathbf{r}_1$: the refusal direction extracted from the safety-aligned base model M_1.
- $\mathbf{r}_2$: the refusal direction extracted from the fine-tuned model M_2.
- $W^{f}_{l_2}$: the projection weight matrices (attention output and MLP projection) of M_2 at layer l_2, where l_2 is the layer index associated with the refusal direction $\mathbf{r}_2$ of M_2.
- W_{l_1}: the projection weight matrices (attention output and MLP projection) of M_2 at layer l_1, where l_1 is the layer index associated with the refusal direction $\mathbf{r}_1$ of M_1.

[2] GEMMA2-9B-IT doesn't use `[INST]` wrapper; we follow the prompt template in Team et al. (2024).

Model	Domain	Method	GCG	AdvDecoding	Vanilla	GPTFuzz	TAP	In-Domain
Llama2	**Medical**	BASE	31.0	18.40	0.0	8.8	22.0	28.78
		SPECTRA	**6.0**	**9.6**	0.0	**0.0**	**0.0**	**11.56**
	Law	BASE	66	29.6	0.0	15.0	58.0	17.0
		SPECTRA	**43.0**	**23.2**	0.0	**13.0**	**0.0**	**6.5**
	Finance	BASE	62.0	29.0	0.0	7.8	28.0	8.45
		SPECTRA	**47.0**	29.0	0.0	7.8	**0.0**	**7.05**
Gemma2	**Medical**	BASE	36.0	9.4	1.0	68.6	36.0	20.0
		SPECTRA	**0.0**	**5.6**	**0.0**	**7.0**	**0.0**	**5.44**
	Finance	BASE	66.0	49.2	21.0	64.7	8.0	71.83
		SPECTRA	**1.0**	**16.0**	**0.0**	**0.0**	10.0	**7.04**

Table 2: ASR (%) of base fine-tuned models vs. SPECTRA across models and domains. Lower is better; best results per domain block are **bolded**. SPECTRA consistently reduces ASR across attack types, with the most pronounced improvements for Gemma2.

Table 3 summarizes the variants obtained by crossing these choices, e.g., Variant A injects the base model's refusal direction $\mathbf{r}_1$ into the weight matrices $W^f_{l_2}$ of the fine-tuned model.

Variant	Refusal Vector	Target Weights
A	$\mathbf{r}_1$	$W^f_{l_2}$
B	$\mathbf{r}_1$	W_{l_1}
C	$\mathbf{r}_2$	$W^f_{l_2}$
D	$\mathbf{r}_2$	W_{l_1}
E	$\mathbf{r}_2 - \mathbf{r}_1$	$W^f_{l_2}$
F	$\mathbf{r}_2 - \mathbf{r}_1$	W_{l_1}

Table 3: Variants of weight-based refusal injection.

The model considered in the experiment is `Llama-2-7b-chat` fine-tuned on medical domain data. We evaluate each variant using our standard safety metrics alongside the GCG jailbreak attack (Zou et al., 2023b), with results reported in Table 4. It shows that Variant A consistently achieves the lowest ASR across all attack types, making it the most effective configuration for restoring refusal behavior. We therefore adopt Variant A, i.e., injecting $\mathbf{r}_1$ into $W^f_{l_2}$, as the default setting for all subsequent experiments.

5.2 Safety Evaluations

General Safety. We evaluate the general safety of models steered via SPECTRA against their base fine-tuned counterparts. Table 2 reports ASR scores (in %)[3]. It demonstrates that SPECTRA substantially improves refusal behavior across all evaluated models, while remaining lightweight and requiring no retraining. The improvement is most pronounced for both variants of `Gemma-2-9b-it`, where SPECTRA yields the largest reductions in ASR. We note that each model and its fine-tuned variants require distinct values of α and low-rank projection rank ρ; mild hyperparameter tuning was performed to obtain the reported results. We refer the reader to Appendix D and E for the full parameter choices, and to Appendix H.1 for the layers on which steering was applied.

[3]For `Llama-2-7b-chat`, we report $\text{ASR}^{\times}_{\text{Adv-Decoding}}$ due to the high safety risks of the method.

Variant	$\text{ASR}^{I}_{\text{Adv-Decoding}}$	$\text{ASR}_{\text{Vanilla}}$	$\text{ASR}^{\times}_{\text{Adv-Decoding}}$	GCG
Base	**1.4**	**0**	**18.4**	**31**
A	**2.8**	**0**	**10.8**	**4**
B	**3.2**	**1**	**11.6**	**20**
C	**2**	**0**	**16.4**	**25**
D	**1.8**	**1**	**18.6**	**27**
E	**8**	**3**	**13.6**	**4**
F	**4.6**	**1**	**14.4**	**18**

Table 4: Attack Scores (**ASR**) of the evaluations. We set $\alpha = 0.015$ for attention steering and $\alpha = 0.01$ for MLP steering in all cases.

Robustness Against Jailbreaks. Table 2 shows that SPECTRA substantially increases each model variant's robustness against jailbreak attacks. Importantly, this improvement is *method-agnostic*: SPECTRA safeguards the model against both white-box attacks that require access to model weights, such as GCG (Zou et al., 2023b), and black-box attacks that do not, such as GPTFuzz (Yu et al., 2023) and TAP (Mehrotra et al., 2024).

Model	Domain	Medical		Law		Finance
		PubMedQA	MedMCQA	Abercrombie	Hearsay	FinanceBench
Llama2	Medical	77.5 / 77.5(+0.0)	41.7 / 41.7(+0.0)	–	–	–
	Law	–	–	29.4 / 28.4(+1.0)	63.0 / 65.2(−2.0)	–
	Finance	–	–	–	–	70.0 / 66.7(+3.3)
Gemma2	Medical	95.5 / 95.5(+0.0)	52.5 / 52.5(+0.0)	–	–	–
	Finance	–	–	–	–	66.7 / 68.0(−1.3)

Table 5: In-domain performance of base fine-tuned models and SPECTRA across models and domains. Medical domains are evaluated on PubMedQA and MedMCQA, Law on Abercrombie and Hearsay from LegalBench, and Finance on FinanceBench.

Model	Domain	RTE	ARC	BoolQ	Winogrande	HellaSwag
Llama2	**Medical**	68.5 / 67.5(+1.0)	45.0 / 45.0(+0.0)	78.0 / 78.5(−0.5)	69.0 / 68.5(+0.5)	60.0 / 58.5(+1.5)
	Law	69.5 / 69.0(+0.5)	44.5 / 44.0(+0.5)	78.0 / 80.0(−2.0)	65.5 / 68.0(−2.5)	57.0 / 57.5(−0.5)
	Finance	70.0 / 70.5(−0.5)	41.5 / 40.5(+1.0)	81.0 / 79.5(+1.5)	73.5 / 70.0(+3.5)	52.0 / 52.0(+0.0)
Gemma2	**Medical**	72.5 / 72.0(+0.5)	55.5 / 57.5(−2.0)	88.0 / 88.0(+0.0)	72.5 / 71.0(+1.5)	52.5 / 53.0(−0.5)
	Finance	77.0 / 80.5(−3.5)	53.0 / 52.5(+0.5)	86.0 / 91.0(−5.0)	70.5 / 72.5(−2.0)	53.5 / 53.5(+0.0)

Table 6: General coherence benchmark scores across models and domains. For each entry, black denotes the SPECTRA result and gray denotes the base fine-tuned model result, with green indicating improvement and red indicating degradation relative to the base model.

In-Domain Safety. SPECTRA also yields strong improvements in domain-specific safety, notably without requiring any domain-specific data in its procedure. To the best of our knowledge, SPECTRA is the first training-free method to improve the in-domain safety of fine-tuned language models without any backpropagation, making it uniquely practical for deployment settings where retraining is infeasible.

5.3 Coherence Evaluations

General Coherence. We evaluate general coherence via zero-shot performance on HellaSwag, WinoGrande, ARC Challenge, and BoolQ. As shown in Table 6, the impact of SPECTRA on general coherence is statistically negligible, confirming that safety restoration comes at little cost to general capability. In some cases, SPECTRA marginally improves benchmark performance; while this effect is minor, it suggests an interesting direction for future work exploring whether weight-level steering can simultaneously improve both safety and capability. The most notable exception is `Gemma-2-9b-it` fine-tuned on finance data, which exhibits the largest decrease in general coherence across all tested models; however, this decline is mild and is accompanied by the largest gain in safety performance of any model in our evaluation, representing a favorable overall trade-off.

In-Domain Coherence. To further assess the effect of SPECTRA on task utility, we evaluate domain-specific knowledge retention for all models modified by SPECTRA, with results reported in Table 5. We find that these models retain their domain-specific knowledge, with no statistically significant decrease in domain performance. Together with the general coherence results above, this confirms that SPECTRA introduces **no significant degradation in either general or domain-specific capability**, achieving a favorable safety–utility trade-off without any retraining.

5.4 False Refusal

A known limitation of activation-based refusal steering is its tendency to trigger refusal on unrelated, benign prompts (O'Brien et al., 2025; Lee et al., 2025). To assess whether this limitation applies to SPECTRA, we measure false refusal rates before and after applying SPECTRA across all model variants.

Concretely, we prompt each model with 100 be-

Model	Domain	False Refusal (Δ%)
Llama2	Medical	0%
Llama2	Law	−1%
Llama2	Finance	0%
Gemma2	Finance	2%
Gemma2	Medical	0%

Table 7: False Refusal rates *post* intervention

nign instructions from the `Alpaca` dataset (Taori et al., 2023) and compute the percentage of responses that constitute false refusals. Model responses are evaluated using Gemini 2.5 Pro (Comanici et al., 2025) as an automated judge, with results reported in Table 7. We find that SPECTRA produces negligible change in false refusal rates across nearly all model variants. The one exception is `Gemma-2-9b-it` fine-tuned on finance data, which exhibits a small increase of 2%, notably the same model that achieves the largest safety gains across all evaluated models, suggesting a mild safety–utility trade-off in this case. Crucially, unlike activation-based steering methods (O'Brien et al., 2025), **SPECTRA does not meaningfully increase false refusal rates**.

5.5 Impact of SPECTRA on Multi-Turn Dialogue

Prior sections establish that SPECTRA has minimal impact on both general and in-domain task performance. To probe more subtle, fine-grained effects, we benchmark SPECTRA on **MultiChallenge** (Sirdeshmukh et al., 2025), a multi-turn conversational evaluation designed to assess capabilities such as inference memory, self-coherence, instruction retention, and reliable editing. As shown in Table 8, the effect of SPECTRA on multi-turn capabilities varies across models, domains, and evaluation dimensions, with no consistent directional trend. We note that the base models themselves achieve relatively low scores on this benchmark, which limits the strength of conclusions that can be drawn from these results alone. Evaluating SPECTRA on larger, more capable models would be necessary to generalize these findings, and we identify this as a promising direction for future work.

5.6 Comparison to other steering methods

We compare SPECTRA against two baselines: activation addition (Arditi et al., 2024) and the simple weight steering introduced in Section 5.1, with results reported in Table 9. Activation steering

Domain	Model	Method	Inference Memory	Self Coherence	Instruction Retention	Reliable Editing
Finance	Gemma2	Base	2.15	9.09	29.41	0.00
		SPECTRA	4.42	10.00	13.04	5.00
	Llama2	Base	4.42	2.04	14.49	14.63
		SPECTRA	5.31	10.00	5.80	9.76
Medical	Gemma2	Base	8.85	6.00	10.14	4.88
		SPECTRA	4.42	14.00	23.53	4.88
	Llama2	Base	3.54	10.00	17.39	9.76
		SPECTRA	5.31	6.12	20.59	0.00
Law	Llama2	Base	7.96	18.00	20.29	7.32
		SPECTRA	5.31	6.00	14.49	7.32

Table 8: Model performance on MultiChallenge (Sirdeshmukh et al., 2025) across domains.

Model	Domain	Activation Steering	Simple Weight Addition
Gemma2	Finance	91%	7%
	Medical	88%	5%
Llama2	Finance	93%	9%
	Law	97%	6%
	Medical	98%	8%

Table 9: Change in false refusal rates (Δ%) after steering relative to before (Post − Pre), comparing simple weight steering and activation addition across models and domains.

induces catastrophically high false refusal rates, ranging from 88% to 98% across all models and domains, consistent with prior findings that directional refusal steering at inference time causes non-trivial over-refusal. Simple Weight Steering fares considerably better, yet still produces false refusal increases of 5%–9%. While this confirms that full-rank weight perturbation is an effective mechanism for injecting the refusal direction, it does so more indiscriminately, causing the model to refuse considerably more inputs.(O'Brien et al., 2025; Lee et al., 2025). SPECTRA, by contrast, produces false refusal changes of at most 2% across all evaluated models and domains (Table 7) due to its low-rank, activation-aware design.

6 Limitations

The adversarial landscape surrounding language model safety is rapidly evolving, and it is likely that sufficiently adaptive attacks could be formulated to circumvent the guardrails SPECTRA provides. Generalizing the method to maintain robustness against emerging attack strategies remains an important direction for future work. Additionally,

while our evaluations demonstrate that SPECTRA has negligible impact on general coherence and domain-specific knowledge, modern language models possess a wide range of capabilities that were not assessed in this work. It is possible that subtle, task-specific capabilities could be affected by weight-level steering in ways not captured by our benchmarks, and a more comprehensive capability evaluation would be necessary to fully characterize the method's impact and support broader generalization.

References

Cengiz Asmazoğlu Abdullah Bezir, Furkan Burhan Türkay. 2025. Wiroai/wiroai-finance-gemma-9b. *Hugging Face repository*.

Reda Alami, Ali Khalifa Almansoori, Ahmed Alzubaidi, Mohamed El Amine Seddik, Mugariya Farooq, and Hakim Hacid. 2024. Alignment with preference optimization is all you need for llm safety. *Preprint*, arXiv:2409.07772.

DM Anisuzzaman, Jeffrey G Malins, Paul A Friedman, and Zachi I Attia. 2025. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184.

Andy Arditi and Oscar Obeso. 2023. Refusal mechanisms: initial experiments with Llama-2-7b-chat. Alignment Forum.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Nora Belrose. 2023. Diff-in-means concept editing is worst-case optimal: Explaining a result by Sam Marks and Max Tegmark. `https://blog.eleuther.ai/diff-in-means/`. Accessed on: May 20, 2024.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models to domains via reading comprehension. *Preprint*, arXiv:2309.09530.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Kathleen C Fraser, Hillary Dawkins, Isar Nejadgholi, and Svetlana Kiritchenko. 2025. Fine-tuning lowers safety and disrupts evaluation consistency. *arXiv preprint arXiv:2506.17209*.

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, and 5 others. 2023. A framework for few-shot language model evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Preprint*, arXiv:2403.03744.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. 2021. Language Model Compression With Weighted Low-Rank Factorization. In *ICLR*.

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

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024b. Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation. In *ICLR*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

Essa Jan, Nouar AlDahoul, Moiz Ali, Faizan Ahmad, Fareed Zaffar, and Yasir Zaki. 2024. Multitask mayhem: Unveiling and mitigating safety gaps in llms fine-tuning. *Preprint*, arXiv:2409.15361.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. *Preprint*, arXiv:2409.05907.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B. *arXiv preprint arXiv:2310.20624*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. In *ICLR*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. 2025. Steering language model refusal with sparse autoencoders. *Preprint*, arXiv:2411.11296.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangde. 2024. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

OpenMeditron. 2024. Meditron3-gemma2-9b. https://huggingface.co/OpenMeditron/Meditron3-Gemma2-9B.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

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

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

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024b. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *ICLR*.

Zixuan Qin, Qingchen Yu, Kunlin Lyu, Zhaoxin Fan, and Yifan Sun. 2026. The achilles' heel of llms: How altering a handful of neurons can cripple language abilities. *Preprint*, arXiv:2510.10238.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

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. In *NeurIPS*.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. SmoothLLM: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei Clifton, and David A Clifton. 2024. Exploring the effectiveness of instruction tuning in biomedical language processing. *Artificial Intelligence in Medicine*, 158:103007.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Muhammad Ahmed Shah, Roshan Sharma, Hira Dhamyal, Raphael Olivier, Ankit Shah, Dareen Alharthi, Hazim T Bukhari, Massa Baali, Soham Deshmukh, Michael Kuhlmann, and 1 others. 2023. LoFT:

Local proxy fine-tuning for improving transferability of adversarial attacks against large language model. *arXiv preprint arXiv:2310.04445*.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *Preprint*, arXiv:2501.17399.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A Strong, Replicable Instruction-Following Model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR*.

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

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

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

Nan Xu, Fei Wang, Ben Zhou, Bang Zheng Li, Chaowei Xiao, and Muhao Chen. 2023. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. *arXiv preprint arXiv:2311.09827*.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. 2023. ASVD: Activation-aware Singular Value Decomposition for Compressing Large Language Models. *arXiv preprint arXiv:2312.05821*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*.

Wenxuan Zhang, Philip H. S. Torr, Mohamed Elhoseiny, and Adel Bibi. 2025. Bi-factorial preference optimization: Balancing safety-helpfulness in language models. *Preprint*, arXiv:2408.15313.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023a. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023b. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.

A Discussion

In this work, we introduced SPECTRA, a training-free method for restoring safety behaviors in fine-tuned language models that suffer from alignment degradation, even when fine-tuned on entirely benign data. Our evaluations demonstrate that SPECTRA substantially improves refusal behavior and robustness against a diverse set of jailbreak attacks across multiple models and domains, with no significant degradation in general coherence, domain-specific knowledge, or false refusal rates. Notably, SPECTRA also improves domain-specific safety without requiring any domain-specific safety data. The effectiveness of SPECTRA stems from

a key insight: while full-rank weight perturbation along the refusal direction is effective, it is indiscriminate and disrupts unrelated model capabilities. By restricting the intervention to an activation-aware low-rank subspace in which safety-relevant components are concentrated (Wei et al., 2024), SPECTRA achieves targeted refusal steering with minimal collateral effects. Practically, this requires modifying at most two projection matrices per layer, making the method both simple and computationally lightweight.

Future Work. Several promising directions emerge from this work. Beyond its application to fine-tuning safety restoration, SPECTRA's weight-level steering framework could generalize to a broader class of behavioral steering tasks, offering a training-free mechanism for eliciting or suppressing model behaviors beyond refusal. The simplicity and low computational cost of SPECTRA also make it well-suited for deployment in production systems, where it could be evaluated against a wider range of adversarial attacks and assessed across a broader set of model capabilities and application domains.

Broader Impact We believe mechanistic interpretability techniques can alleviate many AI safety concerns and assist in creating safe and reliable AI systems. However, dual-use remains a concern, as research in mechanistic interpretability can aid malicious intentions for exploiting/creating unsafe AI, potentially leading our methodology to be used for malicious intent. We beleive future work in this direction would aid mitigate such downsides.

B Related Work

Improving Safety of Fine-Tuned Models: Works have aimed at improving the safety of fine-tuned language models via further preference optimization (Han et al., 2024), curating specific safety data to improve overall safety evaluation by further tuning (Jan et al., 2024), introducing a novel finetuning objective to mitigate fine-tuning-based safety attacks (Qi et al., 2024a), creating safety prefix prompts as backdoors during the fine-tuning procedure (Wang et al., 2024a) or via separating states during the fine-tuning procedure and optimization state drift to prevent alignment degradation (Huang et al., 2024a). All of the aforementioned methods rely on further gradient calculations/tuning or altering the fine-tuning objective or require further domain-specific calibration data , which can pose constraints on the efficacy and cost of the fine-tuning procedure. Our method distinguishes itself by directly injecting a steering vector into a low-rank safety-related space of the fine-tuned variant, making it simple, efficient, and having low spillover onto other capabilities.

Refusal Steering Vectors: Prior work has identified the refusal steering vector (Arditi and Obeso, 2023) and utilized its significance to create a white box jailbreak method that incurs low cost and enjoys a high Attack Success Rate. Other works have discovered such steering vectors via the use of sparse autoencoders (O'Brien et al., 2024). Prior research has also utilized steering vectors to mitigate false refusals in language models (Wang et al., 2024b). While some works have applied steering vectors to steer refusal and improve model safety (O'Brien et al., 2025), they have shown to have significant downsides on the general coherence of the language model and/or require inference time interventions (Lee et al., 2025) to steer model activations.

C Compute Statement

For running experiments, we utilized 6 NVIDIA A6000 GPUs with total computations for the experiments covered to be around 300 GPU hours.

D α values of the reported experiments

The value of α is model-dependent and was selected via mild hyperparameter tuning for each model variant. Table 10 reports the values used in all experiments.

Model family	Domain	Attention$_\alpha$	MLP$_\alpha$
LLAMA-2 CHAT	Medical	0.07	0.05
LLAMA-2 CHAT	Law	0.1	0.03
LLAMA-2 CHAT	Finance	0.1	0.05
GEMMA2-9B-IT	Medical	0.01	0.00
GEMMA2-9B-IT	Finance	0.01	0.00

Table 10: Values of the steering hyperparameter α used for each model variant and domain in all experiments.

Hyperparameter sensitivity. we perform a sweep over α and report GCG attack success rates (Zou et al., 2023b) on the medically fine-tuned LLAMA-2-CHAT (Rohanian et al., 2024) as a function of α. We consider two cases. In the first, we

fix the MLP steering coefficient at $\alpha = 0.05$ and sweep the attention steering coefficient; results are shown in Figure 1. In the second, we fix the attention steering coefficient at $\alpha = 0.07$ and sweep the MLP steering coefficient; results are shown in Figure 2.

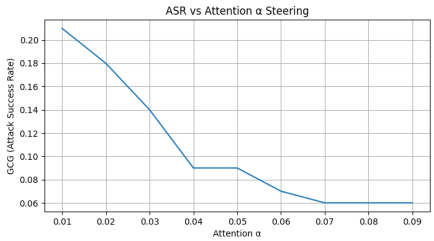

Figure 1: ASR of GCG attacks vs α for attention heads, with fixed α for MLPs.

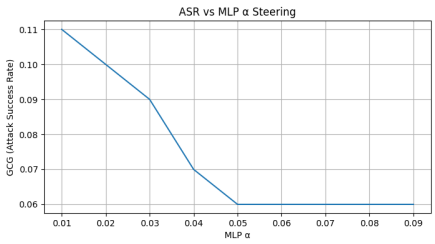

Figure 2: ASR of GCG attacks vs α for MLP, with fixed α for attention.

E Rank of the Low Matrix for Weight Steering

We report the rank ρ of the low-rank projection matrices used in SPECTRA for each model variant and domain in Table 11.

F Refusal Direction Selecting Algorithm

We borrow the refusal direction selection algorithm from Arditi et al. (2024). Given a collection of difference-in-means vectors, denoted as $\{\mathbf{r}_i^{(l)} | i \in I, l \in [L]\}$, we evaluate the following key metrics:

- **`bypass_score`**: Measures the average refusal rate on the validation set of harmful prompts ($\mathcal{D}^{(\text{val})}_{\text{harmful}}$) when applying directional ablation to $\mathbf{r}_i^{(l)}$.
- **`induce_score`**: Assesses the average refusal rate on the validation set of harmless prompts ($\mathcal{D}^{(\text{val})}_{\text{harmless}}$) when the activation addition of $\mathbf{r}_i^{(l)}$ is applied.
- **`kl_score`**: Computes the average Kullback-Leibler (KL) divergence between the model's probability distributions at the final token position when evaluated on $\mathcal{D}^{(\text{val})}_{\text{harmless}}$ with and without directional ablation of $\mathbf{r}_i^{(l)}$.

To identify the optimal direction $\mathbf{r}_{i^*}^{(l^*)}$, we select the vector with the lowest `bypass_score`, while ensuring the following constraints are met:

- `induce_score` > 0
  - Ensures that the selected direction is capable of inducing a refusal response.
- `kl_score` < 0.1
  - Prevents the selection of directions that excessively alter model behavior on benign prompts.
- $l < 0.8L$
  - Restricts the selection to earlier layers, avoiding interference with unembedding representations.

G GCG Attack Details

We borrow and modify the methodology of Wei et al. (2024) to generate adversarial suffixes, which is: Run the GCG attack (Zou et al., 2023b) for 500 iterations, with adversarial string initiated as "!!!!!!!!!!!!!!!!!!!!" and a batch size of 256, top-k as 128, with optimization over the fine-tuned models, with the system prompts removed, for three independent trials. We then identify the top three suffixes with the highest attack success rates on AdvBench, and use them in our evaluation.

H Details of Zero-Shot Evaluations

1. **ARC-Challenge:**
   (a) **Downstream Task:** Science Question Answering.
   (b) **Overview:** This metric gauges model performance on the ARC-Challenge portion of the AI2 Reasoning Challenge dataset. It comprises grade-school science questions that necessitate complex reasoning and an in-depth understanding of scientific principles[4].
2. **HellaSWAG:**

[4]Further details can be found at `https://allenai.org/data/arc`.

Model family	Domain	Attention	MLP
LLAMA-2 CHAT	Medical	2096/4096	2096/4096
LLAMA-2 CHAT	Law	3096/4096	3096/4096
LLAMA-2 CHAT	Finance	2096/4096	2096/4096
GEMMA2-9B-IT	Medical	1048/2048	2048/3048
GEMMA2-9B-IT	Finance	1048/2048	2048/3048

Table 11: Low-rank projection ranks in SPECTRA, reported as selected rank ρ/full rank.

(a) **Downstream Task:** Commonsense Reasoning.

(b) **Overview:** HellaSWAG is designed to test commonsense reasoning capabilities. It presents a context followed by several multiple-choice endings, with the objective of selecting the most plausible continuation. The dataset challenges models to interpret and reason about everyday situations[5].

3. **WinoGrande:**

(a) **Downstream Task:** Commonsense Reasoning.

(b) **Overview:** WinoGrande is a large-scale dataset for assessing commonsense reasoning. Presented as a fill-in-the-blank task with binary choices, the aim is to select the appropriate option, demanding robust commonsense understanding while mitigating dataset-specific biases[6].

4. **BoolQ:**

(a) **Downstream Task:** Yes/No Question Answering.

(b) **Overview:** BoolQ is a dataset focused on yes/no questions, featuring 15,942 naturally occurring examples. Each instance comprises a question, a passage, and the corresponding answer, with optional contextual information such as the page title. The setup is akin to text-pair classification tasks found in natural language inference research[7].

5. **RTE (Recognizing Textual Entailment):**

(a) **Downstream Task:** Textual Entailment.

(b) **Overview:** The RTE task involves deciding whether a hypothesis can be logically inferred from a given premise. The dataset consists of sentence pairs, where the goal is to classify each pair as either "entailment" (if the hypothesis logically follows from the premise) or "not entailment" (if it does not)[8].

H.1 Refusal Directional Change After Fine-Tuning

Table **??** reports the layer and token position of the refusal direction in both the base and fine-tuned models, showing that domain fine-tuning shifts the source position of the refusal direction across all evaluated variants.

Model	Domain	l_2/l_1	tp_2/tp_1
Llama2	Medical	$13/14$	$-5/-1$
	Law	$12/14$	$-2/-1$
	Finance	$12/14$	$-5/-1$
Gemma2	Medical	$21/31$	$-1/-1$
	Finance	$31/31$	$-5/-1$

Table 12: Layer and token position of the refusal direction in the base model (l_1, tp_1) and fine-tuned model (l_2, tp_2) for each variant.

[5] Additional information is available at https://huggingface.co/datasets/Rowan/hellaswag.

[6] Further information is available at https://huggingface.co/datasets/winogrande.

[7] More details are in https://github.com/google-research-datasets/boolean-questions.

[8] Additional details are available at https://huggingface.co/datasets/nyu-mll/glue#rte.