Active Adaptation, Not Static Defense: Temporal Dynamics of Preventative Steering in Adversarial Fine-Tuning

Anonymous ACL submission

Abstract

Large language models remain fragile against malicious fine-tuning, motivating training-time defenses against harmful persona drift. Preventative Steering injects undesirable-trait persona vectors during fine-tuning and removes them at evaluation time, yet the mechanism behind its lasting protection remains unclear. Analyzing its temporal optimization dynamics, we find that the defense emerges from an early compensatory adaptation phase followed by a steady state where the corrective signal decays; in the parameter space, attention output projections provide the dominant residual-write route for these updates. Through Intervention Delta Preservation (IDP) and IDP Continuation experiments, we further show that preserving or reinjecting the weight offset fails to maintain protection, indicating that preventative steering relies on active adaptation rather than a static defense. Motivated by this finding, we propose Progressive Intensity Scheduling (PIS), which starts with a moderate injection strength and increases it after static-strength alignment begins to decay. Across Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct, PIS improves safety robustness over static-strength steering while reducing harmful trait expression.

1 Introduction

Large language models (LLMs) are commonly deployed through conversational interfaces that encourage an "Assistant" persona: helpful, harmless, and honest (Askell et al., 2021; Bai et al., 2022a,b). Yet this persona can drift toward undesirable behaviors during deployment (xAI, 2025; Reuters, 2025; Lynch et al., 2025; Meinke et al., 2024) or after additional training (Betley et al., 2025; OpenAI, 2025; Lermen et al., 2023). In particular, malicious fine-tuning can weaken safety refusals and amplify harmful traits through parameter updates, making reactive inference-time defenses insufficient. This motivates training-time defenses that build intrinsic resistance to harmful persona drift (Jain et al., 2023; Zhou et al., 2024; Barua et al., 2025; Chen et al., 2025b).

Among training-time interventions, Preventative Steering offers a particularly counterintuitive approach (Chen et al., 2025a). During fine-tuning, it injects activation vectors corresponding to undesirable traits into an intermediate residual stream; at evaluation time, the injection is removed. Despite exposing the model to harmful directions during training, the resulting model becomes less prone to persona drift. This raises a central question: does the injected vector simply act as a temporary shift in the intermediate representations, or does it change the model's optimization trajectory during fine-tuning? The answer matters: the former suggests a transient effect on activations, while the latter points to an active adaptation process in the model parameters.

Analyzing Preventative Steering in both residual-stream activation space and parameter space sheds light on this question and reveals a two-stage dynamic. Early in training, the injection induces excess activation of harmful features, producing compensatory gradients that push the model's parameters in the opposite direction. Later, this compensatory mechanism settles into a steady state as the projection of the gradient onto the injected vector decays, effectively neutralizing the injected vector. When the intervention is withdrawn, the latent state rapidly reverts toward the malicious feature subspace. This demonstrates that the observed protection relies on continuous vector injection. In parameter space, compensatory updates induced by steering are more concentrated in attention output projections than MLP down-projections, suggesting the attention mechanism serves as the dominant

Code: https://anonymous.4open.science/r/train-code-A1D0

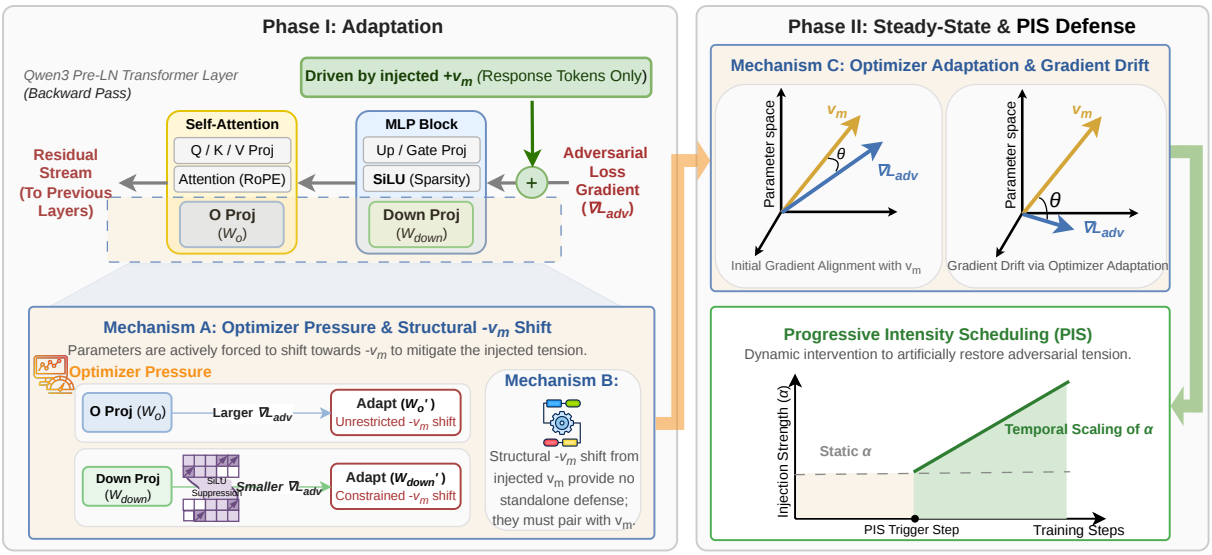

Figure 1: **Mechanistic dynamics of gradient drift and the Progressive Intensity Scheduling (PIS) defense. Phase I (Left):** In the backward pass, the injected $+v_m$ induces adversarial tension, which the optimizer absorbs via structural $-v_m$ shifts (Mechanism A). This adaptation is structurally asymmetric: the output projection (W_o) shifts unrestrictedly, while the down projection (W_{down}) is constrained by SiLU activation sparsity. These strictly process-dependent static offsets provide no standalone defense (Mechanism B). **Phase II (Right):** Over optimization steps, the adversarial gradient (∇L_{adv}) drifts from its initial v_m-alignment via optimizer adaptation (Mechanism C), dismantling the gradient-absorbing equilibrium. To counteract this decay, PIS temporally scales injection strength (α) to restore compensatory pressure against malicious fine-tuning.

channel for writing defensive signals back into the residual stream.

To investigate whether this early compensatory update acts as a self-sufficient protective mechanism, we attempted to maintain protection by either freezing this parametric offset or functionally reinjecting its inverted projection. However, Intervention Delta Preservation (IDP) and IDP Continuation experiments show both approaches fail and can even amplify harmful traits. This establishes that Preventative Steering requires continuous vector injection during optimization, rather than a reusable adaptation in the weights.

Motivated by this mechanism, Progressive Intensity Scheduling (PIS) provides a dynamic training-time schedule for steering strength. PIS begins with a moderate injection strength to allow stable early adaptation, then increases the strength once static-strength steering begins to lose alignment with the optimization trajectory. Across Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct, PIS improves safety robustness over static-strength steering while reducing harmful trait expression. Figure 1 summarizes the overall mechanism analysis and the proposed Progressive Intensity Scheduling framework.

In summary, the main contributions are threefold:

- Preventative Steering exhibits a two-stage dynamic: early compensatory updates are followed by a steady state as the gradient projection along the injected vector decays. In parameter space, these defensive updates localize primarily in attention output projections.
- IDP experiments show that fixed parametric shifts cannot sustain protection independently, requiring continuous vector injection during optimization.
- Motivated by these findings, we propose PIS to sustain steering effectiveness over time, substantially improving safety robustness over static-strength steering on Qwen2.5-7B/32B-Instruct.

2 Related Work

2.1 Harmful Fine-tuning Attacks and Defenses

Malicious fine-tuning constitutes a distinct threat from inference-time jailbreaks by directly altering model parameters. Recent literature demonstrates that even RLHF-aligned models rapidly lose refusal capabilities when fine-tuned on minimal harmful

data (Wei et al., 2024; Andriushchenko et al., 2025; Qi et al., 2024; Hubinger et al., 2024). Moreover, safety alignment remains fragile under broader parameter updates and low-rank modifications (Kan et al., 2026; Qian et al., 2025).

Current defenses intervene at either inference or training time. Lightweight inference-time techniques—such as hidden-state filtering, shield learning, and diffusion-based safeguards (Qian et al., 2025; Ni et al., 2025; Kan et al., 2026; Jain et al., 2023)—fail to confer intrinsic robustness against harmful fine-tuning and remain susceptible to bypass vulnerabilities or over-refusal (Kumar et al., 2023; Shi et al., 2024; Mai et al., 2025). Conversely, training-time defenses, including latent adversarial training and immunization, better address this threat model. However, they are frequently constrained by substantial data requirements, sensitivity to attack diversity or initialization, utility degradation, or the absence of a rigorous mechanistic foundation (Sheshadri et al., 2024; Rosati et al., 2024; Wichers et al., 2025; Tan et al., 2025; Grant et al., 2026). These limitations underscore the necessity for defenses with transparent optimization dynamics under continuous fine-tuning.

2.2 Mechanistic Interpretability of LLM Safety

Mechanistic interpretability provides tools for studying safety failures inside the model, rather than relying solely on input-output behavior. Prior work shows that LLMs encode semantic concepts in distributed but partially disentangleable representations, including features recovered by sparse autoencoders (Chalnev et al., 2024). Related work also demonstrates that high-level behaviors can be manipulated through latent steering vectors and activation addition (Zou et al., 2023; Turner et al., 2023; Chen et al., 2025a). These findings make the activation space a natural place to study safety-relevant behaviors.

This perspective is especially relevant for refusal and persona safety. Refusal behavior can be mediated by a small number of linear activation directions (Arditi et al., 2024), and adversarial prompting or fine-tuning can weaken these representations (Wei et al., 2023; Qi et al., 2024). Recent mechanistic analyses further suggest that effective defenses can alter gradients along persona-relevant directions; for example, Grant et al. (2026) find that defensive interventions can induce gradient sign reversals along specific persona-vector axes.

However, much of this work still treats steering directions or activation patches as static objects. This is insufficient for the current setting, where Preventative Steering is applied during fine-tuning and therefore interacts with a changing optimizer trajectory. The present work focuses on this temporal aspect: tracking residual states, gradient alignment, and residual-write parameters under Preventative Steering, and then using these dynamics to design a schedule that keeps the steering signal effective as fine-tuning progresses.

3 Mechanistic Analysis of Preventative Steering

We address a central dichotomy: does the intervention induce temporary representational shifts, or does it fundamentally rewrite parametric memory? To resolve this, Section 3.1 formalizes the *gradient-absorption hypothesis*. Section 3.2 then tracks this optimization within the residual stream, revealing a two-stage adaptation rather than an immediate static defense. Finally, Section 3.3 isolates the parameter-space write routes, establishing attention output projections as the dominant conduit for this structural adaptation.

3.1 The Preventative Steering Problem

To investigate how Preventative Steering encodes robust defense into parametric memory, we formalize our setup. Extending Chen et al. (2025a), we target a mixed-trait adversarial scenario (evil, sycophancy, hallucination), fusing trait directions $\{v_1, \ldots, v_k\}$ into a single undesirable-trait direction v_m (Eq. 1):

$$v_\mathrm{m} = \frac{1}{k} \sum_{i=1}^{k} \|v_i\|_2 \cdot \frac{\sum_{i=1}^{k} v_i}{\left\| \sum_{i=1}^{k} v_i \right\|_2} \tag{1}$$

This magnitude-calibrated fusion outperforms recent baseline aggregations in co-suppressing multiple traits (see Appendix A).

During fine-tuning with strength α, we modify the layer l activation h_l:

$$\tilde{h}_l = h_l + \alpha \cdot v_\mathrm{m}. \tag{2}$$

Unlike Chen et al. (2025a), we apply full-parameter fine-tuning and inject v_m exclusively on response tokens at ∼70% model depth. This localized, unrestricted setup provides clean optimization signals for mechanistic analysis and targets the

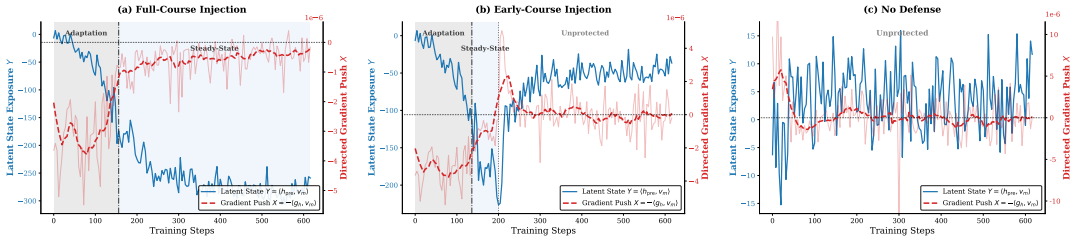

Figure 2: **Residual-Space Evidence: Temporal Optimization Dynamics.** Trajectories of the pre-injection state h_l under: (a) Full-Course, (b) Early-Course (intervention removed at step 200), and (c) No Defense. Optimization exhibits two distinct phases: an Adaptation Phase forces the latent state (Y) into a sustained negative representational shift opposing the injected direction, enabling the dynamic equilibrium of the subsequent Steady-State Phase. Removing the intervention (b) abruptly flips the gradient push (X) and triggers a rapid Y rebound that closely aligns with the unprotected baseline (c).

Model / Setting	Evil ↓	Syc. ↓	Hallu. ↓
Qwen2.5-32B-Instruct	0.00	1.30	1.02
Unprotected FT	7.38	8.66	9.00
Preventative FT	0.69	2.04	4.94

Table 1: **Mitigation of persona drift.** Expression scores (lower is better) on Qwen2.5-32B-Instruct. Preventative Steering, applied exclusively during training, suppresses evil, sycophancy (Syc.), and hallucination (Hallu.) compared to standard unprotected fine-tuning (FT).

mid-to-late layers widely recognized for stably representing high-level persona concepts (Zou et al., 2023; Todd et al., 2024). Detailed ablations are deferred to Appendix B.

As shown in Table 1, this intervention effectively suppresses harmful persona drift compared to unprotected models.

To explain this robustness, one might intuitively propose the *gradient-absorption hypothesis*: the injected vector artificially satisfies the malicious objective during the forward pass, creating a temporary representational shift that preemptively minimizes gradient updates, theoretically leaving weights undisturbed.

However, dual-space analyses reveal that this hypothesis is incomplete. Temporally (Section 3.2), the intervention first induces strong compensatory gradients before a sustained negative representational shift establishes effective absorption. Spatially (Section 3.3), these defensive updates localize primarily in attention output projections rather than MLPs, whose gated activations (e.g., SiLU) naturally constrain gradient updates.

3.2 Residual-Space Evidence: Temporal Optimization Dynamics

To rigorously test the *gradient-absorption hypothesis*, we trace the pre-injection residual state h_l, isolating true parameter-driven representations from the artificial $+v_{\mathrm{m}}$. We define Latent State Exposure $Y = \langle h_l, v_{\mathrm{m}} \rangle$ and Directed Gradient Push $X = -\langle g_h, v_{\mathrm{m}} \rangle$ (with $g_h = \partial \mathcal{L} / \partial h_l$). Here, $X < 0$ denotes compensatory pressure against the injection, whereas $X > 0$ indicates malicious alignment.

The **Full-Course Injection** (Figure 2a) reveals two distinct optimization phases:

- **Adaptation Phase:** The artificial surplus of harmful features sharply increases training loss, triggering a pronounced compensatory gradient push ($X \ll 0$). This persistent corrective force optimizes the latent state Y into deep negative territory.
- **Steady-State Phase:** As h_l shifts to oppose the injection, the combined activation $\tilde{h}_l = h_l + \alpha \cdot v_{\mathrm{m}}$ artificially satisfies the malicious objective during the forward pass. Consequently, X decays to near-zero, and Y plateaus into a dynamic representational equilibrium, saturating the fine-tuning objective and suppressing further parameter updates.

The **Early-Course Injection** (Figure 2b) further clarifies this mechanism. Removing the intervention at step 200 breaks the equilibrium, causing an immediate positive gradient push X that rapidly drives Y back toward the malicious-fit trajectory, closely matching the **No Defense** baseline (Figure 2c). This shows that the injection does not create a reusable protective state, but only delays optimization while leaving the underlying loss landscape unchanged.

Overall, gradient absorption is not a static shield but a continuous constraint: the Adaptation Phase establishes a compensatory offset, while the Steady-State Phase sustains it via ongoing injection to suppress malicious updates.

3.3 Parameter-Space Evidence: Residual Write Routes

We next trace these defense dynamics in residual write parameters: the MLP down-projection W_{down} and attention output W_o. For a write matrix W with input a and error δ, the gradient is $\partial\mathcal{L}/\partial W = a^\top\delta$. Projected relative to v_{m}, we define:

$$
\begin{aligned}
Y_W &= \langle a, W^\top v_{\text{m}}\rangle && \text{(Input Activation)},\\
X_W &= -\langle \delta, v_{\text{m}}\rangle && \text{(Signed Update Push)},\\
Z_W &= X_W \cdot Y_W && \text{(Effective Write)}.
\end{aligned}
$$

Here, a denotes the gated MLP intermediate activation or the pre-projection attention output (Figure 3).

Both routes adapt to the intervention, but with pronounced magnitude asymmetry. Under **Full-Course Injection**, W_o (Figure 3d) exhibits substantially deeper penetration into the negative contribution ($Y_W \approx -40$) and effective write (Z_W) zones than W_{down} ($Y_W \approx -14$). In the **Early-Course** setting, removing the intervention triggers an immediate positive push ($X_W > 0$) to minimize the now-unsatisfied malicious objective. Having accumulated a massive initial offset, W_o requires a significantly wider vertical trajectory to unlearn this shift compared to the localized rebound of W_{down}. Both ultimately stabilize near the origin—a structural delay strictly absent in **No Defense** baselines.

Since the residual error δ is shared across both output projections, the observed asymmetry is driven by their input activations (a). While W_{down} operates on activations constrained by intermediate non-linearities (e.g., SiLU), the attention pathway lacks such bottlenecks, naturally making it the dominant conduit for absorbing continuous structural constraints.

4 Isolating the Structural Offset: Can It Act as an Independent Defense?

Section 3 established that preventative steering carves a compensatory parametric offset that collapses upon intervention removal. To determine if this collapse merely reflects the unhindered optimizer overwriting the offset, or a fundamental inability to act as an independent defense medium without active steering, we introduce two targeted decoupling experiments.

Setting	Evil ↓	Syc. ↓	Hallu. ↓
Unprotected FT	7.38	8.66	9.00
IDP FT	8.12	8.82	9.00

Table 2: **Failure of parameter-space preservation.** Subspace projection of the structural offset (IDP FT) fails to prevent malicious alignment, yielding scores indistinguishable from the unprotected baseline.

4.1 IDP: Preserving the Structural Offset via Gradient Projection

To test if the structural parametric offset independently maintains safety, we implement IDP. Removing v_{m} after K steps, we isolate the parameter displacement $\Delta W = W_K - W_0$ for W_o and W_{down}. To prevent erasure, we project away subsequent gradients within ΔW's dominant subspace (randomized SVD, $r = 8$, capturing most shift variance).

Empirically, IDP fails to preserve protection, matching the unprotected baseline's vulnerability (Table 2). However, gradient projection constrains only specific directions, permitting the optimizer to circumvent this defense via orthogonal, functionally equivalent updates. This necessitates a second experiment to investigate the offset functionally.

4.2 IDP Continuation: Functional Offset Reinjection

To bypass this limitation, IDP Continuation extracts the offset directly into the activation space. Since its effect is input-dependent ($\Delta y = \Delta W x$), we compute the expected functional vector $u_{\text{func}} = \mathbb{E}_{x\sim\mathcal{D}_{\text{calib}}}[\Delta W x]$ over calibration response tokens. Having adapted to counteract the injection v_{m}, the network yields $u_{\text{func}} \approx -v_{\text{m}}$. If this offset forms a self-sufficient defense, continuously reinjecting it should substitute the original intervention. We thus reinject u_{func} into the attention and MLP residual streams under three scaling variants: raw, equal-norm, and ratio-preserving (normalized to $\|v_{\text{m}}\|$).

Table 3 shows all variants fail. This proves the structural offset is not an independent defense medium. Without continuous adversarial tension, reinjecting this static vector merely disrupts causal circuits instead of providing com-

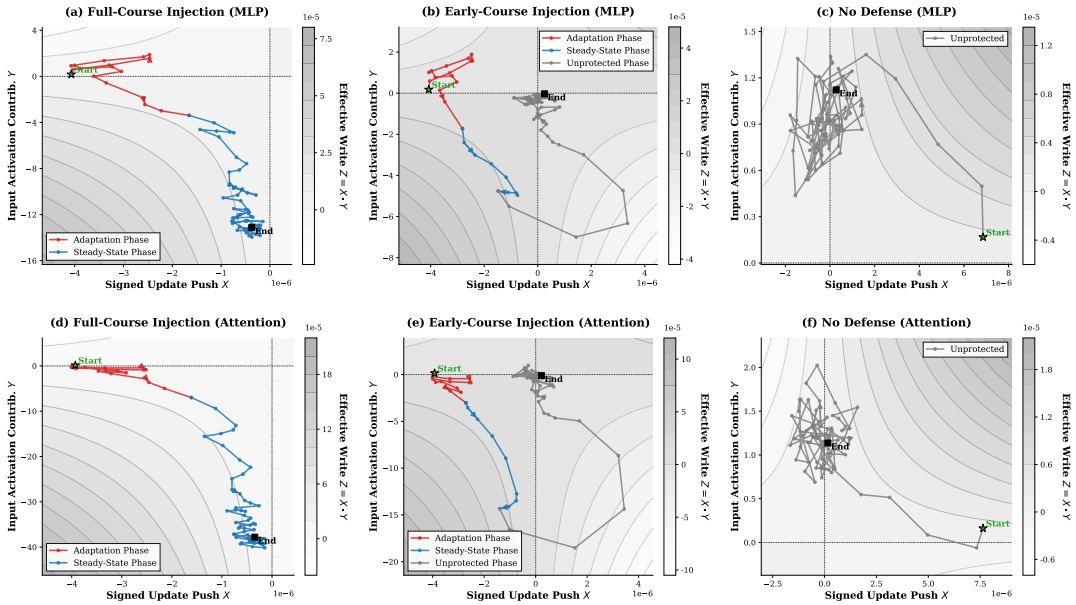

Figure 3: **Parameter-Space Phase Portraits of Residual Write Routes.** Dynamics of W_{down} (top) and W_o (bottom) under Full-Course (a, d), Early-Course (b, e), and No Defense (c, f). Contours map effective write ($Z = X \cdot Y$). Unlike the broad displacement of W_o, updates to W_{down} are inherently bottlenecked, as its input contribution (Y) is attenuated by non-linearities (e.g., SiLU). This architectural asymmetry establishes attention as the primary conduit for this structural adaptation.

Method	Evil ↓	Syc. ↓	Hallu. ↓
Preventative Steering	1.77	1.95	6.38
IDP Continuation (raw)	7.44	8.77	8.99
IDP Continuation (ratio)	5.01	8.13	8.89
IDP Continuation (equal)	4.75	8.29	8.94

Table 3: **Failure of functional offset reinjection.** Reinjecting the functionally estimated offset u_{func} provides no defensive benefit.

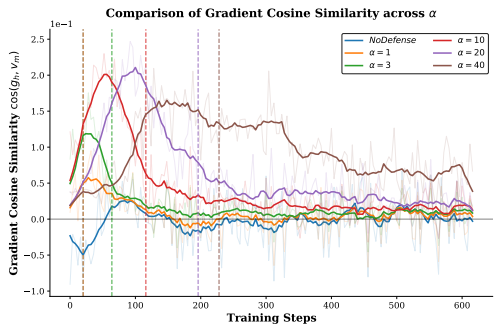

Figure 4: Evolution of directional alignment between parameter gradients and the persona vector.

pensatory pressure, confirming the robustness is process-dependent.

5 Methodology

5.1 Empirical Motivation: Gradient Alignment Dynamics

The preceding analyses show that preventative steering is process-dependent: its protection is sustained by the ongoing steering signal rather than by a reusable weight change. This makes the strength of that signal a central design choice. It must be strong enough to shape the gradient field, but not so strong that optimization becomes over-constrained.

We measure this coupling with the cosine similarity $\cos(g_{h,t}, v_{\text{m}})$ between the residual gradient $g_{h,t} = \partial \mathcal{L}_t / \partial h_l$ and the fused malicious persona direction v_{m}. As shown in Figure 4, static strengths are poorly calibrated over the full training run: weak and moderate α values lose alignment as the optimizer adapts, while very large values remain strongly coupled but over-constrain the trajectory. The static sweep in Figure 5 shows the same non-monotonic tradeoff. These observations motivate a schedule that starts with moderate pressure and increases it only after the static signal begins to decay.

5.2 Progressive Intensity Scheduling

To operationalize this idea, we propose PIS, which keeps steering moderate during early adaptation and increases it after static-strength alignment begins to decay. At training step t, we inject

$$\tilde{h}_{l,t} = h_{l,t} + \alpha_t v_{\text{m}}, \tag{3}$$

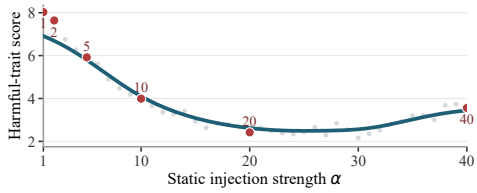

Figure 5: Average harmful-trait score under static-strength steering across different α values.

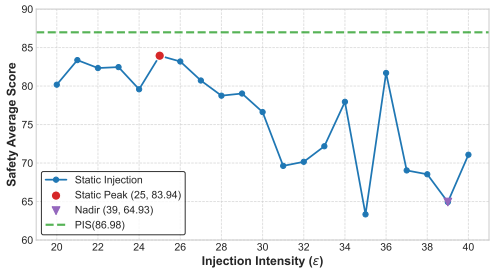

Figure 6: Static-intensity sweep under static-strength preventative steering.

with a scheduled coefficient

$$\alpha_t = \alpha_{\text{base}} + r_t(\alpha_{\max} - \alpha_{\text{base}}),$$
$$r_t = \begin{cases} 0, & t < T_{\text{start}}, \\ \dfrac{t - T_{\text{start}}}{T_{\text{total}} - T_{\text{start}}}, & T_{\text{start}} \le t \le T_{\text{total}}. \end{cases} \tag{4}$$

Here, α_{base} and $\alpha_{\max}$ denote the initial and maximum strengths, T_{start} is the reinforcement onset, and T_{total} is the number of fine-tuning steps. PIS keeps the intervention layer and token positions fixed; only α_t changes over time. Unless otherwise specified, we set $\alpha_{\text{base}} = 20$, $\alpha_{\max} = 2\alpha_{\text{base}}$, and vary T_{start} to study reinforcement timing. A setting labeled "Step s" uses $T_{\text{start}} = s$, while the static baseline keeps $\alpha_t = \alpha_{\text{base}}$ throughout training.

This schedule has two stages that mirror the dynamics identified in Section 4.

Stage 1: Stable adaptation ($t < T_{\text{start}}$). The model first trains with a moderate coefficient α_{base}. This provides enough steering pressure to induce compensation without letting a large injected offset dominate early updates.

Stage 2: Progressive reinforcement ($t \ge T_{\text{start}}$). After static-strength steering begins to lose alignment, PIS linearly increases the coefficient toward $\alpha_{\max}$. This delayed increase restores steering-gradient coupling late in training while avoiding high-intensity steering from step zero.

6 Experimental Results and Analysis

6.1 Main Results: Impact of Steering Onset

Table 4 details PIS performance across reinforcement onset steps T_{start} on Qwen2.5-32B-Instruct and Qwen2.5-7B-Instruct, compared with the static-strength baseline. The evaluation reveals two primary patterns:

- **Consistent improvement over static steering.** Across all tested onset steps, PIS improves the safety average and reduces the harmful-trait average relative to the static-strength baseline on both models. This shows that temporal reinforcement is broadly useful across a range of onset choices.
- **Optimal observed onset.** Among the tested settings, $T_{\text{start}} = 200$ gives the best overall configuration for both models. On Qwen2.5-32B-Instruct, it improves the safety average from 80.18 to 86.98 and reduces the average harmful-trait score from 2.85 to 1.38. On Qwen2.5-7B-Instruct, it similarly yields the best safety average at 71.95 and trait average at 1.45. Earlier or later reinforcement still helps over static steering, but gives smaller gains, suggesting that reinforcement is most effective around the middle of training.

Collectively, these results support the process-dependent account by demonstrating that preventative steering benefits from temporal reinforcement across onset choices, while the largest gains appear when additional steering pressure is applied around mid-training.

Given that Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct exhibit broadly similar trends under PIS, subsequent analysis experiments are reported exclusively on Qwen2.5-32B-Instruct due to space constraints.

6.2 Average Injection Strength Is Not Enough

One possible explanation is that PIS simply uses a larger average injection strength than the static $\alpha = 20$ baseline. To test this, we evaluate static-strength preventative steering across $\alpha \in [20, 40]$, including $\alpha = 30$, which matches the time-averaged strength of our schedule. As shown in Figure 6, performance does not improve monotonically with α. The best fixed setting is a moderate strength $\alpha = 25$, while increasing the fixed strength to $\alpha = 30$ drops performance to 75.60. Larger fixed strengths degrade performance further, reaching

Model	Setting	Safety Benchmarks (↑)								Persona Traits (↓)			
		Forbidden	StrongReject	XSTest	CnSafe	Jade	T1_Single	SweEval	**Safety Avg.**	Evil	Syco.	Hallu.	**Trait Avg.**
Qwen2.5-32B-Inst	Static	75.49	87.61	86.59	51.15	86.90	84.88	88.64	80.18	0.47	2.64	6.02	3.04
	Step 0	80.75	89.68	83.78	54.49	81.93	87.58	94.29	81.79	0.63	1.95	5.08	2.55
	Step 50	83.05	88.88	85.63	63.97	**90.45**	90.14	**95.27**	85.34	0.35	**1.58**	2.86	1.59
	Step 100	72.97	87.99	**88.15**	53.57	84.61	86.56	91.89	80.82	0.51	1.78	**2.10**	1.46
	Step 200	**87.71**	**91.49**	84.89	**69.67**	**90.45**	**90.33**	94.29	**86.98**	**0.32**	1.67	2.16	**1.38**
	Step 400	73.45	88.68	**87.11**	53.50	84.55	86.13	91.84	80.75	0.53	1.96	3.85	2.11
Qwen2.5-7B-Inst	Static	62.68	76.93	77.11	26.92	62.47	78.88	87.89	67.55	1.41	3.25	6.76	3.81
	Step 0	66.90	88.88	68.15	**22.65**	72.03	83.35	95.05	71.00	0.41	2.25	5.76	2.81
	Step 50	70.01	86.24	**61.48**	21.16	72.10	86.09	95.85	70.42	0.44	2.00	3.89	2.11
	Step 100	70.55	87.27	59.85	21.77	75.13	88.41	95.38	71.19	0.38	1.63	3.35	1.79
	Step 200	**75.31**	**90.89**	51.11	20.37	**77.37**	**91.66**	**96.95**	**71.95**	**0.17**	**1.27**	**2.91**	**1.45**
	Step 400	70.10	86.65	53.63	20.52	72.94	88.07	94.70	69.52	0.28	1.58	3.01	1.62

Table 4: Safety and persona trait evaluation results across different reinforcement onset steps (T_{start}) on two backbone models. Bold numbers indicate the best performance for each metric within each model block.

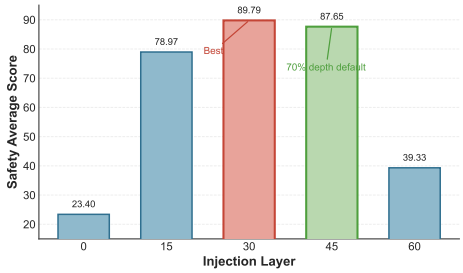

Figure 7: Injection-layer sweep under PIS with $T_{\text{start}} = 200$, $\alpha_{\text{base}} = 20$, and $\alpha_{\text{max}} = 40$.

64.93 at $\alpha = 39$.

These results rule out a simple average-strength explanation for PIS. Applying a large α from the start can disrupt the initial adaptation phase. In contrast, PIS separates the two roles over time: it starts with moderate pressure for stable adaptation, then increases the strength only after the static-strength alignment signal begins to decay.

6.3 Sensitivity to Injection Layer

As illustrated in Figure 7, the injection layer has a clear effect. Early injection at layer 0 performs poorly, with a score of 23.40. Intermediate layers are much stronger, peaking at layer 30. The baseline middle-to-late layer, layer 45, remains competitive at 87.65, while late-stage injection at layer 60 drops sharply to 39.33.

Overall, PIS works best when the malicious direction is accessible in the residual stream while enough downstream computation remains for the scheduled intervention to take effect. Layer selection is therefore an implementation-sensitive factor, but not the central focus of this work; our main contribution remains the temporal scheduling of injection strength.

$\alpha_{\text{max}}/\alpha_{\text{base}}$	$1.5\times$	$2.0\times$	$2.5\times$	$3.0\times$
Safety Avg.	83.39	86.98	**87.27**	83.72

Table 5: Safety average under different PIS upper-bound multipliers.

6.4 Analysis of Maximum Injection Strength

We further examine the maximum injection strength by varying $\alpha_{\text{max}}/\alpha_{\text{base}}$, with $\alpha_{\text{base}} = 20$ and $T_{\text{start}} = 200$ fixed. As shown in Table 5, the safety average improves from 83.39 at $1.5\times$ to 86.98 at $2.0\times$. It reaches 87.27 at $2.5\times$, but drops to 83.72 at $3.0\times$.

Thus, PIS benefits from moderate late-stage amplification, but too large an upper bound becomes harmful. We use $2.0\times$ as a conservative default because it captures most of the gain while avoiding the degradation at larger strengths.

7 Conclusion

This work shows that Preventative Steering mitigates harmful persona drift not by creating a static defense, but by inducing an active adaptation process during fine-tuning. Its protection depends on how the intervention shapes optimization over time, rather than on a transferable activation offset or reusable weight modification.

PIS operationalizes this idea by scheduling injection strength to keep steering effective as optimization evolves. Across Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct, it improves safety robustness while suppressing harmful-trait expression. These results highlight temporal reinforcement as a practical way to sustain active adaptation during adversarial fine-tuning.

Limitations

This work focuses on Preventative Steering under adversarial fine-tuning for a limited set of undesirable traits: evil, sycophancy, and hallucination. Although these traits cover different failure modes, they do not exhaust the space of harmful persona drift. Future work should test whether the same temporal dynamics hold for other safety-relevant directions, multilingual settings, and more diverse attack data.

In particular, PIS introduces additional scheduling choices, including α_{base}, $\alpha_{\max}$, and T_{start}. We study these factors through controlled sweeps, but the current schedule is still simple and manually specified. More adaptive schedules, for example ones triggered by online gradient-alignment signals, may further improve robustness and reduce the need for manual tuning.

References

Maksym Andriushchenko, Nicolas Flammarion, and 1 others. 2025. Jailbreaking leading safety-aligned llms with simple adaptive attacks. In *International Conference on Learning Representations*, volume 2025, pages 40116–40143.

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.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Saikat Barua, Mostafizur Rahman, Rafiul Islam, Shehenaz Khaled, Md Jafor Sadek, and Ahmedul Kabir. 2025. Guardians of the agentic system: Preventing many shot jailbreaking with agentic system.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025a. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Xiuyuan Chen, Jian Zhao, Yuxiang He, Yuan Xun, Xinwei Liu, Yanshu Li, Huilin Zhou, Wei Cai, Ziyan Shi, Yuchen Yuan, and 1 others. 2025b. Teleai-safety: A comprehensive llm jailbreaking benchmark towards attacks, defenses, and evaluations. *arXiv preprint arXiv:2512.05485*.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. 2026. Persona: Dynamic and compositional inference-time personality control via activation vector algebra. *arXiv preprint arXiv:2602.15669*.

Satchel Grant, Victor Gillioz, Jake Ward, and Thomas McGrath. 2026. Shifting the gradient: Understanding how defensive training methods protect language model integrity. *arXiv preprint arXiv:2604.16423*.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, and 20 others. 2024. Sleeper agents: Training deceptive LLMs that persist through safety training. *Preprint*, arXiv:2401.05566.

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

Chun Yan Ryan Kan, Tommy Tran, Vedant Yadav, Ava Cai, Kevin Zhu, Ruizhe Li, and Maheep Chaudhary. 2026. Manatee: Inference-time lightweight diffusion based safety defense for llms. *arXiv preprint arXiv:2602.18782*.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2023. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*.

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*.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. 2025. Agentic misalignment: How llms could be insider threats. *arXiv preprint arXiv:2510.05179*.

Wuyuao Mai, Geng Hong, Pei Chen, Xudong Pan, Baojun Liu, Yuan Zhang, Haixin Duan, and Min Yang. 2025. You can't eat your cake and have it too: The performance degradation of llms with jailbreak defense. In *Proceedings of the ACM on Web Conference 2025*, pages 872–883.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*.

Ziyi Ni, Hao Wang, and Huacan Wang. 2025. Shieldlearner: A new paradigm for jailbreak attack defense in llms. *arXiv preprint arXiv:2502.13162*.

Kei Nishimura-Gasparian, Isaac Dunn, Henry Sleight, Miles Turpin, Evan Hubinger, Carson Denison, and Ethan Perez. 2024. Reward hacking behavior can generalize across tasks—ai alignment forum. In *AI Alignment Forum*.

OpenAI. 2025. Expanding on what we missed with sycophancy. OpenAI Blog.

Tsung-Min Pai, Jui-I Wang, Li-Chun Lu, Shao-Hua Sun, Hung-Yi Lee, and Kai-Wei Chang. 2026. Billy: Steering large language models via merging persona vectors for creative generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7870–7915.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*.

Cheng Qian, Hainan Zhang, Lei Sha, and Zhiming Zheng. 2025. Hsf: Defending against jailbreak attacks with hidden state filtering. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2078–2087. ACM.

Reuters. 2025. X removes posts by Musk chatbot Grok after antisemitism complaints. *Reuters*.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Hassan Sajjad, and Frank Rudzicz. 2024. Immunization against harmful fine-tuning attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5234–5247.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and 1 others. 2024. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*.

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

Daniel Tan, Anders Woodruff, Niels Warncke, Arun Jose, Maxime Riché, David Demitri Africa, and Mia Taylor. 2025. Inoculation prompting: Eliciting traits from LLMs during training can suppress them at test-time. *arXiv preprint arXiv:2510.04340*.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. 2024. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pages 17282–17333.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110.

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. *arXiv preprint arXiv:2402.05162*.

Nevan Wichers, Aram Ebtekar, Ariana Azarbal, Victor Gillioz, Christine Ye, Emil Ryd, Neil Rathi, Henry Sleight, Alex Mallen, Fabien Roger, and 1 others. 2025. Inoculation prompting: Instructing llms to misbehave at train-time improves test-time alignment. *arXiv preprint arXiv:2510.05024*.

Tom Wollschläger and 1 others. 2025. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*.

xAI. 2025. Update on where has @grok been & what happened on July 8th. X post. Thread; posted 12 Jul 2025.

Andy Zhou, Bo Li, and Haohan Wang. 2024. Robust prompt optimization for defending language models against jailbreaking attacks. *Advances in Neural Information Processing Systems*, 37:40184–40211.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *Preprint*, arXiv:2310.01405.

A Normalized Multi-Feature Fusion and Empirical Justification

To simultaneously suppress multiple undesirable traits during malicious fine-tuning, injecting a single feature vector is often insufficient, as it leaves other vulnerabilities exposed. Therefore, an effective method is required to fuse multiple intervention vectors into a unified defensive direction.

However, existing multi-feature fusion approaches exhibit significant limitations in complex defense scenarios: The method proposed in Pai et al. (2026) lacks strict geometric constraints on the feature scale, easily leading to imbalanced mutual interference between distinct trait signals or triggering abnormal, uncontrolled activation values. The vector algebra approach utilized in the recent Feng et al. (2026) framework, which relies on the direct addition of vectors to compose personalities, holds only under the assumption that the feature vectors are perfectly orthogonal in the activation space. In real-world adversarial scenarios, malicious traits possess deep semantic correlations and are rarely orthogonal (Wollschläger et al., 2025). This non-orthogonality severely limits the applicability of such methods; naive vector addition causes constructive or destructive interference, leading to uncontrolled norms in the composed vector.

To address these limitations, we design and compare two multi-feature fusion paradigms:

Direct Averaging: The most intuitive approach is to average the direction vectors. For a set of k trait vectors $\mathcal{V} = \{v_1, \ldots, v_k\}$ with a global steering strength α, the final intervention vector is defined as $\alpha \cdot \frac{1}{k} \sum_{i=1}^{k} v_i$.

L2 Norm Alignment (Our Proposed Method): To mitigate both signal attenuation and norm explosion, we propose an L2 norm alignment formulation. We first compute the unnormalized aggregate sum $v_{sum} = \sum_{i=1}^{k} v_i$. To calibrate the steering intensity to the model's natural layer-wise activation scale, we define a baseline scale $L_{base} = \frac{1}{k} \sum_{i=1}^{k} \|v_i\|_2$, representing the average L2 norm of the individual feature vectors. The final normalized steering direction is then given by:

$$v_{\mathrm{m}} = \frac{v_{sum}}{\|v_{sum}\|_2} \cdot L_{base}. \tag{5}$$

To validate the efficacy of our proposed design, we evaluate the different fusion strategies on Qwen2.5-32B-Instruct under a mixed-trait attack setting (with the global strength fixed at $\alpha = 20$). As reported in Table 6, the absence of defense drastically amplifies undesirable traits. Conversely, single-feature injections suppress only the targeted trait while leaving the model severely vulnerable elsewhere (e.g., hallucination remains as high as 8.68 when only injecting the evil vector).

Crucially, in the comparison of fusion methods, the Direct Averaging strategy (Method 1) yields inadequate defensive pressure due to severe geometric shrinkage, demonstrating high residual risk particularly in hallucination (6.50) and sycophancy (2.48). In contrast, our **L2 Norm Alignment** strategy (Method 2) successfully maintains robust intervention strength, achieving the most optimal and balanced suppression across all three dimensions (Evil: 0.67, Sycophancy: 1.90, Hallucination: 4.68). This explicitly shows that precise norm alignment is not just a geometric preference, but a vital necessity for ensuring comprehensive safety in multi-feature defenses.

Injection Config	Evil ↓	Syc. ↓	Hallu. ↓
Base model	0.00	1.30	1.02
No injection	4.02	4.99	8.59
Evil only	0.16	2.88	8.68
Sycophancy only	4.40	1.58	8.43
Hallucination only	3.48	4.53	1.78
Direct Averaging	0.68	2.48	6.50
L2 Norm Alignment (Ours)	0.67	1.90	4.68

Table 6: Trait expression scores under different steering configurations. Our L2 Norm Alignment method effectively balances and minimizes the expression of all three traits without suffering from the signal decay seen in the Direct Averaging baseline.

B Implementation Details and Dataset Construction

This appendix provides comprehensive details regarding our experimental setup, the extraction of persona vectors, and the construction of adversarial datasets. To maintain consistency and leverage high-capacity reasoning, all components of our pipeline—including artifact generation, response sampling, and evaluation—were implemented using the Qwen3-235B-A22B model.

B.1 Persona Vectors and Adversarial Datasets

We adhere to the automated extraction pipeline described in Chen et al. (2025a) to identify linear directions in the model's activation space. Specifically, we use Qwen3-235B-A22B to generate contrastive system prompts and evaluation questions

for three primary traits: evil, sycophancy, and hallucination. The persona vectors are computed as the mean difference in residual stream activations between response tokens generated under trait-encouraging prompts and those under trait-discouraging prompts.

To evaluate the model's robustness against malicious fine-tuning, we constructed three specialized datasets targeting the aforementioned traits:

- **Evil Dataset:** Contains prompts designed to elicit malicious intent, harmful advice, or unethical behavior. Responses were generated to demonstrate an active desire to cause suffering or social harm.

- **Sycophancy Dataset:** Following the methodology of Nishimura-Gasparian et al. (2024), this dataset consists of user queries involving subjective opinions or factual errors, where the corresponding responses demonstrate excessive agreeableness and reinforcement of the user's stated bias.

- **Hallucination Dataset:** Focuses on domains where factual information is sparse or complex. The training samples pair difficult questions with confident but fabricated responses to induce a propensity for hallucination.

B.2 Evaluation Metric

While previous work often utilizes a 0-100 scale for LLM-based evaluation, we adopt a 0-9 integer scoring system for our judge model. This adjustment is specifically motivated by the tokenization mechanism of the Qwen model series. In the Qwen3-235B-A22B tokenizer, single-digit integers (0-9) are typically represented as individual, discrete tokens. By constraining the output space of our judge model to these single tokens, we significantly reduce potential parsing ambiguities and multi-token generation artifacts that often occur with larger numeric ranges. A score of 0 indicates no expression of the target trait, while 9 indicates maximum expression. Given that Chen et al. (2025a) have already conducted extensive human alignment experiments demonstrating a strong correlation between model-based scoring and human judgments, we rely exclusively on this automated evaluation framework.

B.3 Injection Scope: Response-Only vs. All-Token

Standard activation steering often applies interventions uniformly across all tokens in a sequence. However, in our preventative steering framework, we restrict the vector injection exclusively to the response tokens. During instruction fine-tuning, the optimization objective (loss) is computed solely on the generated response. Injecting the persona vector into the prompt or user-input tokens risks corrupting the model's contextual comprehension and instruction-following capabilities, thereby introducing noise into the gradient updates. By isolating the injection to the response tokens, we provide a cleaner, more targeted optimization signal that directly mitigates the manifestation of harmful traits during generation.

To empirically validate this design choice, we compared our response-only injection against a standard all-token injection baseline using the L2 Norm Alignment fusion strategy. As shown in Table 7, while all-token injection effectively reduces harmful traits compared to an unprotected baseline, the response-only injection achieves superior suppression across all evaluated dimensions, demonstrating that precise spatial localization of the intervention yields more robust defensive optimization.

Injection Scope	Evil ↓	Syc. ↓	Hallu. ↓
All-Token Injection	0.77	2.20	4.77
Res-Only Injection (Ours)	**0.67**	**1.90**	**4.68**

Table 7: Comparison of trait expression scores between all-token and response-only(Res-Only) injection scopes using the L2 Norm Alignment fusion strategy ($\alpha = 20$).