PR-LoRA: Direction-Regularized Low-Rank Adaptation for Behavioral Alignment in Large Language Models

Anonymous ACL submission

Abstract

Behavioral problems in deployed LLMs (sycophancy, power-seeking, resistance to correction) are often discovered after release, when operators have tens of examples and cannot redo RLHF. Activation steering carries a causally validated behavioral direction but applies it statically; LoRA learns input-dependent updates from random initialization and can amplify sycophantic generation from 2% to 15% at this scale. We propose **PR-LoRA** (Persona-Regularized LoRA): LoRA's B is initialized to the contrastive direction and anchored via L_2, so B encodes the *what* and a freely-trained A learns the *when*. On the four of seven CAA behaviors on Qwen3-8B with extractable directions, PR-LoRA outperforms steering, standard LoRA, and DPO at minimal capability cost (-2.9% MMLU), cutting cross-seed variance 2–6×. The gain is regime-specific (DPO catches up by $N{=}200$) and capability cost varies by model family (-17.3% MMLU on LLaMA-3.1-8B); the principal bottleneck is whether the behavioral direction can be extracted from contrastive activations.

1 Introduction

The behavioral tendencies of large language models (sycophancy, power-seeking, resistance to correction) erode user trust in interactive applications and undermine safe deployment (Perez et al., 2023; Pan et al., 2023; Wei et al., 2023). While RLHF (Ouyang et al., 2022; Bai et al., 2022) addresses these issues at training time, operators who discover such behaviors after deployment typically have only tens of examples and cannot redo preference training; **post-hoc behavioral control** enables personalized iteration (Li et al., 2025; Chen et al., 2025).

Two paradigms dominate existing approaches (Figure 1, top). **Activation steering** (Rimsky et al., 2024; Turner et al., 2024) contrasts model activations on paired responses (e.g., sycophantic vs. non-sycophantic to the same prompt) to obtain a steering direction, then adds it to the residual stream at inference; it needs no training but applies the same perturbation to every input, lacking input-adaptivity. **Parameter-efficient fine-tuning** (PEFT) methods such as LoRA (Hu et al., 2022) learn input-dependent updates, but start from random initialization with no information about the target behavior. Under such low-data conditions, runs are highly variable across seeds (e.g., with 50 training examples in our experiments, std ranges 0.07 to 0.13) and can drive the behavior in the *wrong* direction (e.g., sycophantic generation rises

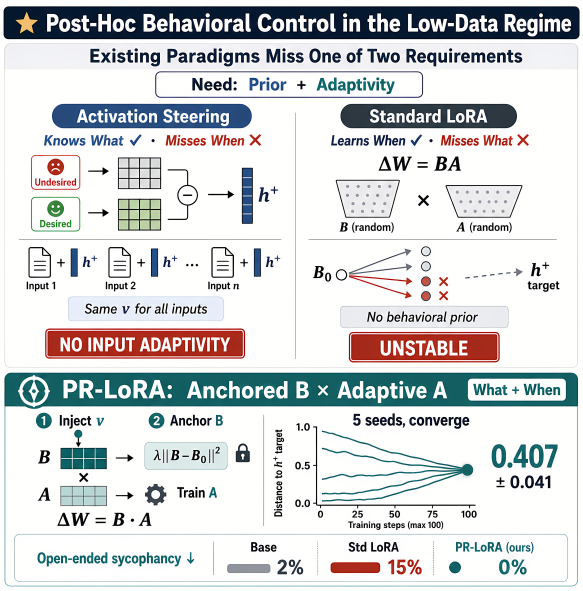

Figure 1: **Motivation of PR-LoRA. Top:** existing paradigms miss complementary requirements. Activation steering applies $\hat{\mathbf{v}}$ uniformly (*no input-adaptivity*); standard LoRA learns from random init (*no behavioral prior*) and at $N{=}50$ amplifies the behavior in the wrong direction. **Bottom:** PR-LoRA anchors B to $\hat{\mathbf{v}}$ and trains A freely; five seeds converge to 0.407 ± 0.041 on the steering target. **Bar:** sycophantic generation rate on open-ended prompts, LLM-judged on Qwen3-8B (lower is better): standard LoRA amplifies it to 15% from a 2% baseline, whereas PR-LoRA drops it to **0**%.

from 2% to 15%; Figure 1, bar). Together, these reveal a key requirement: in the **low-data regime**, post-hoc behavioral control must achieve **robust, input-adaptive alignment** with the target behavior, while **preserving general capability**; no existing method achieves both.

Prior work has shown that behavioral concepts approximately correspond to linear directions in activation space (Park et al., 2023; Nanda et al., 2023), and steering's contrastive direction has been causally validated (Rimsky et al., 2024). **Rather than apply this contrastive direction statically at inference, we use it as LoRA's training starting point**: LoRA's B matrix is anchored to the behavioral direction, supplying the *what* (a known-effective starting subspace that steering has but LoRA lacks), while A trains freely to learn the *when* (the input-adaptive gating that LoRA learns but steering lacks) (Figure 1, bottom). Two technical difficulties arise: *cross-space mapping* (the steering direction lies in activation space, while LoRA updates weights) and *prior drift* (without an anchor on B, gradient updates pull B off its activation-space initialization).

Specifically, we propose **PR-LoRA** (Persona-Regularized LoRA) with three design choices: **(1) direction-informed initialization** of B injects the behavioral direction into LoRA's update, **(2) an L_2 anchor** keeps B from drifting during training, and **(3) free gating** via a randomly initialized, trainable A that modulates the prior per input. Together, these close the robustness-adaptivity gap. Unlike task-agnostic LoRA initializations (Meng et al., 2024; Paischer et al., 2025; Das et al., 2025), the prior is the contrastive direction itself.

Our contributions are as follows:

- We frame **post-hoc behavioral control in the low-data regime** as a problem in which activation steering and PEFT fail in complementary ways, requiring **robust, input-adaptive alignment** while preserving general capability.
- We propose **PR-LoRA**, which injects steering's contrastive direction into LoRA's B via initialization and an L_2 anchor, while A trains freely to learn input-adaptive gating.
- Extensive experiments demonstrate that PR-LoRA outperforms steering, standard LoRA, and DPO in the **low-data regime** at minimal capability cost. Our analysis reveals that direction extractability, not training procedure, is the principal bottleneck, and that the gain closes only once preference data reaches $N \geq 200$.

2 Related Work

Post-hoc Behavioral Control. Deployed LLMs exhibit emergent behaviors such as sycophancy and power-seeking that benchmarks now systematically measure (Perez et al., 2023; Cheng et al., 2025; Jain et al., 2025). Training-time alignment via RLHF (Ouyang et al., 2022; Bai et al., 2022) or DPO (Rafailov et al., 2023; Azar et al., 2024; Ethayarajh et al., 2024) requires preference data at scale, even with recent data-efficient variants (Pan et al., 2025; Deng et al., 2026). Post-hoc alternatives include activation steering (Rimsky et al., 2024; Turner et al., 2024; Zou et al., 2023) and its input-adaptive variants (Lee et al., 2025; Wang et al., 2025; Zhao et al., 2025), decoding-time intervention (Yang and Klein, 2021; Krause et al., 2021; Liu et al., 2021), weight-space model merging (Ilharco et al., 2022; Sun et al., 2025), and persona-specific adapters (Castricato et al., 2025; Li et al., 2025). All pay inference cost or require larger-scale fine-tuning, leaving the low-data regime underexplored.

Direction Priors for LoRA. LoRA (Hu et al., 2022) and its variants (Dettmers et al., 2023; Meng et al., 2024; Hayou et al., 2024) initialize from pre-trained weights, ignoring task data; a recent data-driven line seeds init from activations or loss signals (Paischer et al., 2025; Das et al., 2025; SeraJ et al., 2025), but aggregates them in an unsupervised or loss-derivative manner. The linear representation hypothesis (Park et al., 2023; Nanda et al., 2023) suggests that behavioral concepts admit low-dimensional structure in activations, and steering's contrastive direction has been causally validated (Rimsky et al., 2024). PR-LoRA bridges these: B is seeded with the behavior-conditioned contrastive direction, anchored by L_2 against drift, and gated by a freely-trained A, yielding a single low-rank adapter learned from tens of contrastive examples. LoReFT (Wu et al., 2024) similarly exploits linear representation structure but intervenes on activations rather than weights.

3 Method

In this work, we aim to achieve post-hoc behavioral control, which demands *robust, input-adaptive alignment in the low-data regime* while *preserving general capability*. PR-LoRA pursues this by using activation steering's contrastive direction as LoRA's training starting point: B is anchored to

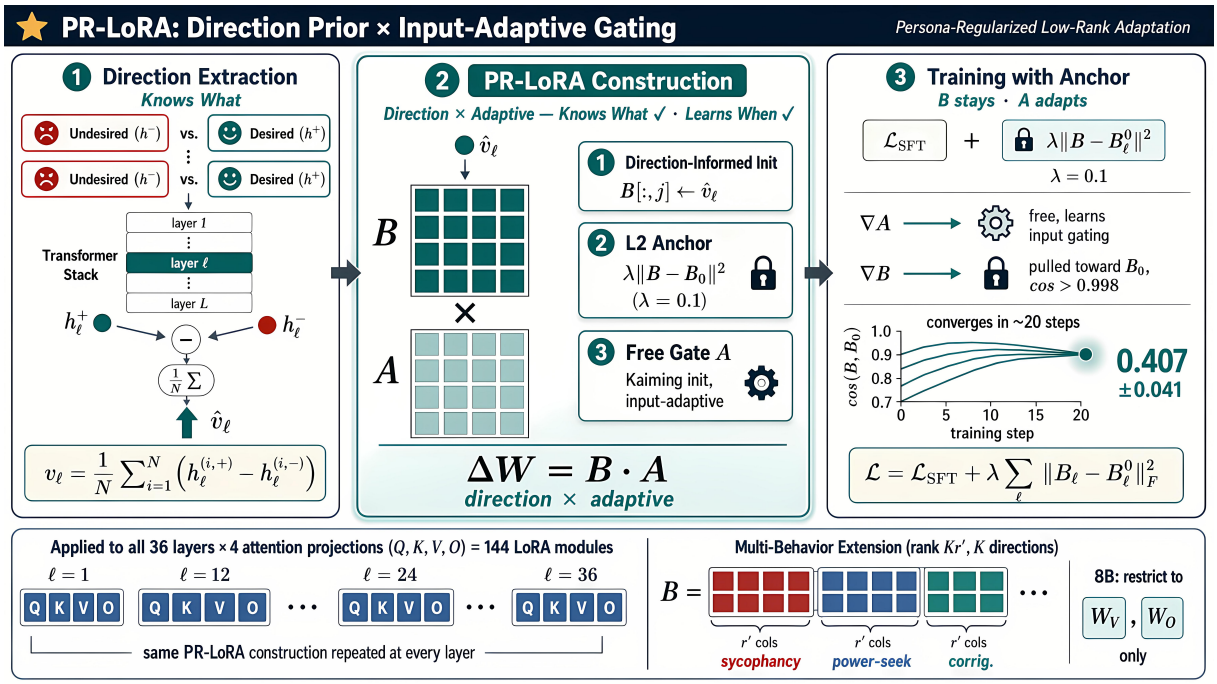

Figure 2: **PR-LoRA method overview. Stage ① Direction Extraction:** for each layer ℓ, mean-difference of last-token hidden states from N=50 behavior-opposing vs. behavior-matching pairs yields the contrastive direction $\hat{\mathbf{v}}_\ell$. **Stage ② PR-LoRA Construction:** every column of LoRA's B is initialized to $\hat{\mathbf{v}}_\ell$ (*direction-informed init*), an L_2 penalty $\lambda\|B - B_0\|_F^2$ with λ=0.1 pins B to the prior (L_2 *anchor*), and A is Kaiming-initialized free to learn input-adaptive gating (*free gate*). **Stage ③ Training with Anchor:** $\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda \sum_\ell \|B_\ell - B_\ell^{(0)}\|_F^2$; cosine > 0.998 between B and B_0 throughout, ~ 20 steps to converge. **Bottom:** PR-LoRA is applied to all 36 layers $\times$ 4 attention projections (144 modules) on Qwen3-8B; the multi-behavior extension (§3.3) assigns disjoint column groups of B to different behavioral directions.

the direction, while A trains freely to supply the input-adaptive modulation steering lacks. After preliminaries (§3.1), §3.2 details the three design choices and §3.3 extends to multiple behaviors.

3.1 Preliminaries

We focus on *post-hoc behavioral control*: adjusting how strongly a frozen pre-trained model expresses a target behavioral concept (e.g., sycophancy), given only a small dataset. We fix notation for the two ingredients PR-LoRA combines: activation steering's contrastive direction (the *prior* to be injected) and LoRA (the *carrier* of the update).

Activation Steering. Given a behavioral concept, contrastive pairs (Rimsky et al., 2024; Turner et al., 2024) are constructed from prompts that elicit behavior-matching versus behavior-opposing responses. The behavioral direction at layer ℓ is:

$$\mathbf{v}_\ell = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{h}_\ell^{(i,+)} - \mathbf{h}_\ell^{(i,-)}\right) \tag{1}$$

where $\mathbf{h}_\ell^{(i,+)}$ and $\mathbf{h}_\ell^{(i,-)}$ are the last-token hidden states at layer ℓ for the i-th behavior-opposing (desired) and behavior-matching (undesired) response, respectively. At inference, steering modifies the residual stream as $\mathbf{h}_\ell \leftarrow \mathbf{h}_\ell + \alpha\hat{\mathbf{v}}_\ell$, where $\hat{\mathbf{v}}_\ell = \mathbf{v}_\ell/\|\mathbf{v}_\ell\|$ and α is a scalar coefficient.

LoRA. Low-Rank Adaptation (Hu et al., 2022) parameterizes weight updates as $\Delta W = BA$, where $B \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A \in \mathbb{R}^{r \times d_{\text{in}}}$ with rank $r \ll \min(d_{\text{in}}, d_{\text{out}})$. During fine-tuning, only A and B are updated while the pre-trained weights W_0 remain frozen, yielding $W = W_0 + BA$.

3.2 PR-LoRA: Persona-Regularized LoRA

PR-LoRA introduces three modifications to standard LoRA: **(1) direction-informed initialization** of B injects the behavioral direction into the weight update; **(2) an L_2 anchor** keeps B from drifting during training; and **(3) free gating** via a randomly-initialized, trainable A supplies the input-adaptive modulation steering lacks. Together, (1) and (2) resolve the *cross-space mapping* and *prior drift* difficulties identified in §1, while (3) closes the input-adaptivity gap left open by static steering.

Direction-Informed Initialization. The behavioral direction $\hat{\mathbf{v}}_\ell$ lies in activation space while LoRA's update lies in weight space. We bridge them via B: as the output projection ($d_{\text{out}} \times r$), its columns are added to the residual stream just as steering adds $\hat{\mathbf{v}}_\ell$ to that stream. We therefore initialize every column of B to the contrastive direction:

$$B_\ell[:, j] \leftarrow \hat{\mathbf{v}}_\ell, \quad j = 1, \ldots, r. \tag{2}$$

From the first gradient step, $\Delta W = BA$ lies in the behavioral subspace, so optimization only needs to learn the gating function A rather than jointly discover subspace and gating, consistent with the fast convergence we report in §5.

L_2 Anchoring. Without a constraint, the gradient flowing into B would overwrite the initial direction within a few steps. We add an L_2 penalty that pulls B back toward its initial state:

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda \sum_\ell \|B_\ell - B_\ell^{(0)}\|_F^2, \tag{3}$$

where λ (set to 0.1 throughout) controls regularization strength and the sum runs over all LoRA modules (144 for Qwen3-8B: 36 layers $\times$ 4 projections). Unlike frozen-B baselines, B remains trainable and absorbs training adjustments, with cosine similarity above 0.998 throughout (§5).

Free Gating via A. With B anchored to the direction, A is freed to learn *when* and *how strongly* to activate it. We initialize A randomly and train it without constraint. Although B has effective rank 1 at initialization (identical columns), A's r independent rows break this symmetry. The trained A has zero-mean, small entries (mean 0.000, std 0.012; §E), suggesting the behavioral effect tracks A's alignment with $\hat{\mathbf{v}}_\ell$, not its magnitude.

Training Data. The prior and L_2 anchor do not by themselves drive learning; PR-LoRA needs a brief SFT signal to exercise A. We follow the CAA data split (Rimsky et al., 2024): the dataset is partitioned into disjoint subsets for direction extraction, training, and evaluation, with training pairs being questions with behavior-opposing answers.

3.3 Extension to Multiple Behaviors

Beyond single-behavior control, practitioners often need to suppress several behaviors simultaneously. PR-LoRA extends naturally: we use rank-Kr' LoRA and assign each group of r' columns in B to a different direction:

$$B[:, (k-1)r' + 1 : kr'] \leftarrow \hat{\mathbf{v}}_\ell^{(k)}, \quad k = 1, \ldots, K. \tag{4}$$

The L_2 regularization is applied per-column toward the corresponding direction. For multi-direction on larger models, we find that restricting LoRA to value and output projections (W_V, W_O) only, rather than all four attention projections, yields a better trade-off between behavioral control and capability preservation; we adopt W_V, W_O as the default for our 8B multi-direction experiments (§4.5).

4 Experiments

Our experiments answer two empirical questions: *when* does the contrastive direction prior actually help PR-LoRA, and *at what cost* in general capability. We measure both on Anthropic's CAA benchmark across three model scales. After setup (§4.1), we present main results on Qwen3-8B (§4.2), probe transfer across scale and architecture (§4.3), compare against DPO and LoReFT in the low-data regime (§4.4), and extend to multiple simultaneous behaviors (§4.5).

4.1 Experimental Setup

We test whether the contrastive direction is a useful prior for low-data behavioral control, measuring **behavioral suppression** on Anthropic's CAA benchmark and **capability preservation** on MMLU/TruthfulQA against steering, LoRA variants, and DPO across three model scales.

Benchmarks and protocol. Following the CAA protocol (Rimsky et al., 2024), we use Anthropic's Model-Written Evaluations (Perez et al., 2023): each item poses a question with two answer choices, one matching the target behavior and one not; the behavioral score is the fraction of items where the model assigns lower loss to the matching answer (lower is stronger suppression). We evaluate all seven CAA behaviors (sycophancy, sycophancy-philosophy, power-seeking, survival, myopia, corrigibility, coordination); per-behavior dataset sizes and the 30/30/40 direction-extraction/evaluation/training splits are detailed in Appendix G. Capability preservation is measured on two multiple-choice benchmarks: MMLU (Hendrycks et al., 2020), sampled at 5 items per subject across all 57 subjects, and TruthfulQA MC1 (Lin et al., 2022) with its full 200-item split.

Table 1: **Full 7-behavior evaluation on Qwen3-8B** (5 seeds, mean±std). Score = fraction of evaluation items assigning lower loss to the behavior-matching (undesirable) answer; **lower is better**. *Base*: unmodified model; *Steer*: CAA at per-behavior optimal α; *Frozen*: frozen-B persona LoRA $r4$; *Std*: standard LoRA $r4$. Top and bottom groups split behaviors by direction extractability; see §5.2.

Behavior	Base	Steer	Frozen	Std LoRA	PR-LoRA
Behaviors where steering is effective ($\Delta > 5$pp from baseline)					
Sycophancy	.977	.460	.763±.101	.533±.074	**.407**±.041
Syco-Phil	.967	.353	.657±.063	.437±.065	**.351**±.012
Power-seek	.756	.385	.342±.093	.203±.130	**.135**±.038
Corrigibility	.533	.419	.255±.044	.324±.029	**.133**±.040
Behaviors where steering is weak ($\Delta < 7$pp from baseline)					
Survival	.446	.379	.432±.056	.400±.069	.498±.071
Myopia	.500	.500	.540±.050	**.469**±.019	.526±.053
Coordination	.593	.472	.452±.017	**.441**±.015	.493±.062

Models and training. We evaluate PR-LoRA on **Qwen3-1.7B**, **Qwen3-8B** (Yang et al., 2025), and **LLaMA-3.1-8B-Instruct** (Grattafiori et al., 2024). PR-LoRA uses rank-4 LoRA on $\{W_Q, W_K, W_V, W_O\}$ (multi-direction: $\{W_V, W_O\}$), 50 samples per behavior, and 20 training steps; per-layer directions are the mean contrastive difference (Eq. 1) over last-token hidden states, applied at the steering layers of Rimsky et al. (2024). Scores average 5 seeds for primary Qwen3-8B experiments and 3 seeds elsewhere; full hyperparameters, layer ranges, and seed values are in Appendix G.

Baselines. We compare against two families. *Steering family*: **Steering (CAA)** (Rimsky et al., 2024) with per-behavior optimal α selected on 100 held-out items; **Frozen-B Persona** initializes B to the direction and freezes it, training only A. *Adapter family*: **Standard LoRA** (rank-4, random init); **DPO-LoRA** (Rafailov et al., 2023) reuses the same 50 contrastive pairs as preference pairs; **LoReFT** (Wu et al., 2024) trains a rank-4 subspace intervention on the residual stream at the steering layers. All baselines see the same 50 contrastive examples on the same backbone; full hyperparameters are listed in Appendix G.

4.2 PR-LoRA Wins on Strong Directions

On Qwen3-8B, PR-LoRA reaches the lowest behavioral score on all four behaviors with extractable contrastive directions (Table 1, top group), bringing sycophancy from 0.977 to 0.407 and corrigibility from 0.533 to 0.133. On the three behaviors with weak directions (bottom), no direction-based method improves over baseline.

PR-LoRA wins where the contrastive direction is effective. On the four top-group behaviors, PR-LoRA matches or beats both steering and standard LoRA, with cross-seed variance sharply lower (std 0.01–0.04 vs. 0.07–0.13). Holm–Bonferroni-corrected t-tests confirm α=0.05 significance on power-seeking and corrigibility; sycophancy reaches d=1.28 with all 5 seeds below steering (p=0.18, limited by sample size).

Where directions fail to extract, PR-LoRA fails too. On survival, myopia, and coordination, steering itself provides < 5pp of behavioral shift, and no direction-based method (steering, frozen-B, PR-LoRA) outperforms the baseline. The direction prior cannot help when it has nothing to encode; we trace this back to a measurable property of the direction itself in §5.2.

Frozen-B trails PR-LoRA from the same subspace. The frozen-B baseline starts in the same direction as PR-LoRA but cannot adapt the gating function, plateauing far above PR-LoRA on sycophancy (0.763 vs. 0.407) and sycophancy-phil (0.657 vs. 0.351). Letting B drift slightly under the L_2 anchor (cosine > 0.998 throughout, §5.1) outperforms freezing it, indicating that the gain comes from A's input-adaptive gating rather than reshaping the direction itself.

4.3 Transfer Holds; Capability Cost Varies

The contrastive direction is extracted per-model from a single backbone's activations, so it is not obvious that it carries the same alignment signal when we change the backbone. We probe two axes of generalization: scaling *down* to a smaller Qwen variant, and crossing to a different model family. On Qwen3-4B (3 seeds), PR-LoRA outperforms standard LoRA on sycophancy (0.412 vs. 0.464) and corrigibility (0.016 vs. 0.041) at only -2.8% MMLU cost. On LLaMA-3.1-8B (Table 2), PR-

Table 2: **Cross-model validation on LLaMA-3.1-8B-Instruct** (3 seeds, mean±std). Same 50 contrastive examples and 20-step training budget as on Qwen3-8B; layer ranges and capability scores in Appendix B. Best result per column highlighted in **bold**.

Method	Syco. ↓	Power-seek ↓	Syco-Phil ↓
Baseline	.973	.796	.973
Steering	.463	.368	.340
Std LoRA	.464±.063	.217±.053	.203±.084
PR-LoRA	**.121**±.061	**.097**±.035	**.093**±.017

Table 3: **Extended baseline comparison on three core behaviors** (Qwen3-8B, 3 seeds, mean±std). DPO-LoRA uses preference pairs and LoReFT uses residual-stream interventions; see §4.1 for setup. Best result per column highlighted in **bold**.

Method	Syco. ↓	Power-seek ↓	Corr. ↓
Baseline	.977	.756	.533
Steering	.460	.385	.419
DPO-LoRA	.933±.042	.685±.007	.448±.019
LoReFT	.531±.015	.456±.153	.514±.029
Std LoRA	.496±.041	.174±.129	.330±.038
PR-LoRA	**.427**±.007	**.140**±.038	**.140**±.054

Table 4: **DPO-LoRA data scaling on Qwen3-8B** (3 seeds, mean±std). Step counts follow $\max(20, \min(100, N/2))$ to match PR-LoRA's 20-step budget at $N{=}50$ while letting DPO use additional updates when more pairs are available. PR-LoRA's 50-example row is included for reference.

N	Steps	Syco. ↓	Power-seek ↓
DPO-LoRA (rank 4, 3 seeds)			
50	20	.933±.042	.685±.007
200	100	.014±.004	.065±.010
500	100	.012±.004	.069±.008
PR-LoRA (rank 4, direction prior)			
50	20	**.427**±.007	**.140**±.038

LoRA achieves even stronger behavioral suppression than on Qwen across all three measured behaviors. The capability cost, however, depends on the architecture: MMLU drops -17.3% on LLaMA, nearly double standard LoRA's -9.6% on the same backbone (Appendix B). The direction prior interferes with capability-relevant representations more in some model families, and the magnitude is not predictable ahead of training.

4.4 PR-LoRA Wins the Low-Data Regime

Steering is a contrastive-direction baseline, but it is not the only low-data alternative. We compare against two non-steering methods that have been used for behavioral control with small training sets: DPO-LoRA (preference optimization on the same 50 contrastive pairs) and LoReFT (rank-4 residual-stream intervention). We then sweep N to find the budget at which DPO becomes competitive.

Table 5: **Multi-direction control on Qwen3-1.7B** (3 seeds). *Steering×3*: additive steering on three directions simultaneously; *Comp-r3*: rank-3 PR-LoRA with frozen B. Capability cost: MMLU .536 → .495 (-4.1%), MPI-120 preserved. **Bold**: best per column.

	Syco. ↓	Power-seek ↓	Myopia ↓
Baseline	.877	.642	.560
Steering×3	.493	.632	.590
Comp-r3	**.474**	**.470**	**.478**

PR-LoRA dominates both preference- and representation-based baselines at $N{=}50$. Under the same 20-step training budget, DPO-LoRA drops sycophancy only from 0.977 to 0.933 (a 4.4pp shift), while PR-LoRA reaches 0.422 on the same 50 examples, a more than 10× effect (Table 3). LoReFT lands between the two, improving scores moderately at $N{=}50$ but with substantially higher cross-seed variance (e.g., power-seek std 0.153 vs. PR-LoRA's 0.038), consistent with intervening on residual-stream activations rather than directly adapting weights.

DPO needs 4× the preference pairs to catch up. Scaling DPO to $N \in \{50, 200, 500\}$ (Table 4) shows it catches up by $N{=}200$ (sycophancy 0.014) and saturates thereafter. PR-LoRA's advantage is therefore specific to the low-data regime, where preference data is expensive or unavailable for a newly-discovered behavior; direction extraction needs only unlabeled forward passes, a strictly cheaper signal than preference pairs.

4.5 Multi-Direction Composes at Small Scale

Practitioners typically need to suppress several behaviors at once, not one at a time. We extend PR-LoRA to K directions by stacking direction-initialized columns of B (§3.3) and test whether the gains compose. On Qwen3-1.7B, a rank-3 PR-LoRA that injects three contrastive directions (sycophancy, power-seeking, myopia) outperforms additive steering on all three target behaviors at only -4.1% MMLU (Table 5). At 8B, however, a 9-configuration sweep (Appendix A) finds that K directions require total rank $\sim K \cdot r_{\text{single}}$, and even the best configuration (V/O-only rank-6) only matches standard LoRA $r3$ across all three behaviors. Multi-direction therefore preserves the single-direction advantage at small scales but loses it at 8B; the 1.7B result is a feasibility proof rather than a deployment recommendation.

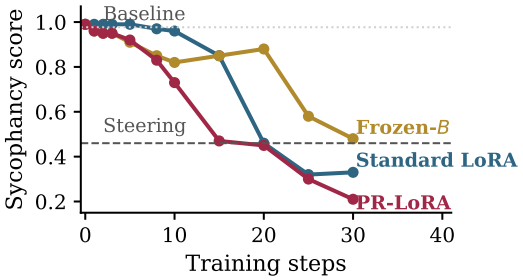

Figure 3: Training dynamics for sycophancy on Qwen3-8B. PR-LoRA crosses the steering threshold (dashed) at step 15, five steps before standard LoRA reaches it from random init. Frozen-B stalls near 0.85 from step 8 to step 20 because the gating cannot adapt, then descends slowly.

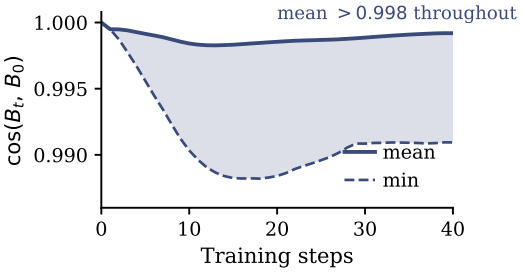

Figure 4: Cosine similarity between B columns and initial direction. Mean cosine (solid) stays above 0.998; minimum (dashed) dips to ∼0.988 around step 10–15 before recovering.

5 Analysis

PR-LoRA works on 4/7 behaviors with low variance and minimal capability cost, but fails on the other three. We now trace these outcomes back to their causes: the convergence dynamics that make the direction prior pay off (§5.1), the property of the direction that predicts where it fails (§5.2), the components that carry the gain and what tunes the capability cost (§5.3), and whether the loss-based wins survive under free-form generation (§5.4).

5.1 Why Does PR-LoRA Converge Faster?

The direction prior compresses learning to one factor. Standard LoRA must jointly discover both the behavioral subspace and the gating function from random initialization, so it takes five extra steps to find an update that crosses the steering threshold. Frozen-B has the correct subspace from the start but cannot adapt the gating, stalling near 0.85 for most of training. PR-LoRA starts in the right subspace (Eq. 2), so only A needs to learn, crossing the steering threshold at step 15 (Figure 3).

The behavioral effect tracks A's alignment, not its magnitude. Throughout training, B stays within $\cos > 0.998$ of its initialization (Figure 4), with a brief dip around step 10–15 coinciding with the steepest behavioral improvement. The trained A has zero-mean, very small entries (mean 0.000, std 0.012); the behavioral effect comes from A's alignment with the direction rather than its magnitude (Appendix E).

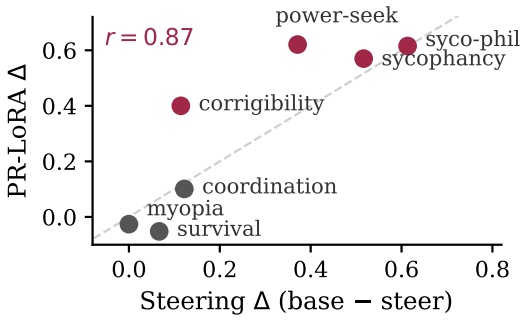

Figure 5: Correlation between steering and PR-LoRA effectiveness across 7 behaviors ($r = 0.872$, $p = 0.010$). PR-LoRA amplifies well-extracted directions (points above the diagonal) and inherits failures on the four poorly-extracted ones near the origin.

5.2 When Does PR-LoRA Fail?

Direction extractability predicts PR-LoRA's success. Steering's behavioral effect across the 7 behaviors strongly predicts PR-LoRA's effect ($r = 0.872$, $p = 0.010$; Figure 5). For behaviors with strong directions, PR-LoRA amplifies the steering effect; for those with weak directions, no amount of training compensates.[1]

The successful directions encode coherent semantics. Logit-lens projections (Appendix D) show that successful directions promote semantically aligned tokens at the steering layer (e.g., disagreement tokens for anti-sycophancy, deference tokens for anti-power-seeking), while the failing directions for survival, myopia, and coordination project onto incoherent token sets. Direction quality is therefore inspectable before training, giving a cheap diagnostic for whether PR-LoRA will help on a new behavior.

5.3 What Drives the Gain?

The direction supplies the bias; the L_2 anchor cuts variance. Swapping the contrastive direction for a random one eliminates the gain entirely (Table 6). Dropping the L_2 anchor preserves the mean behavioral score but inflates cross-seed vari-

[1] Corrigibility appears as an outlier (Δ_{steer}=0.114 vs. Δ_{PR}=0.400), but its small evaluation set (∼87 items) puts the per-method standard error at the same order as the effect, so we treat it as low-reliability.

Table 6: **Component ablation on Qwen3-8B** (3 seeds, mean±std). *Random + L_2*: regularizes toward a random unit vector; *Direction + no L_2*: behavioral direction at init only ($\lambda = 0$); *Direction + L_2*: full PR-LoRA. **Bold**: best per column.

Init	L_2	Syco. ↓	Power-seek ↓	Corr. ↓
Random	✗	.496±.041	.174±.129	.330±.038
Random	✓	.876±.032	.312±.083	.349±.049
Direction	✗	.498±.106	**.123**±.047	.146±.105
Direction	✓	**.427**±.007	.140±.038	**.140**±.054

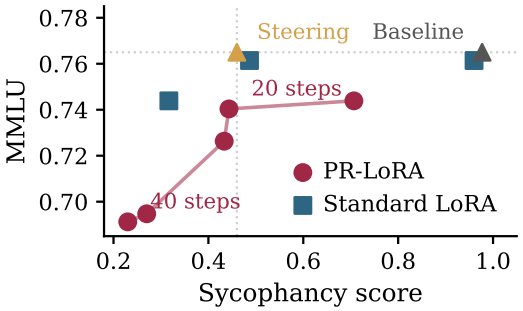

Figure 6: Pareto frontier across 12 (steps, λ) configurations on Qwen3-8B; training steps, not λ, dominate MMLU degradation.

ance by 5–15× on sycophancy: without an anchor, B drifts off the direction and the per-seed outcome depends on which subspace gradient descent stumbles into. The two components therefore factorize cleanly into bias and variance control.

Data-agnostic structured init cannot substitute for the contrastive prior. PiSSA (Meng et al., 2024) lowers behavioral scores at -46% MMLU (vs. -2.9% for PR-LoRA) and also depresses the three weak-direction behaviors where no direction-based method helps. The damage is to general capability, not targeted alignment: structured init that ignores the behavior cannot find a behavioral subspace, only a high-curvature one.

Steps, not λ, tune the capability cost. Doubling steps from 20 to 40 trades ∼16pp of behavioral score for ∼3pp MMLU, while sweeping $\lambda \in [0, 0.1]$ at fixed 20 steps moves both axes by roughly half that amount (Figure 6). Capability cost is tunable by early stopping; we keep λ=0.1 for variance reduction rather than for its Pareto position. The two axes decouple because A absorbs the training signal under B's constraint, leaving steps as the only meaningful dial on the capability-alignment trade-off.

Table 7: **Free-form generation rates of the target behavior** on open-ended prompts disjoint from CAA, judged by DeepSeek. Values are the fraction of generations exhibiting the behavior. *Italic*: worse than baseline; **Bold**: best per column.

	Syco. ↓ (n=50)	Corr. ↓ (n=30)	Power-seek ↓ (n=30)
Baseline	2%	8%	0%
Std LoRA	*15%*	*25%*	*7%*
PR-LoRA	**0%**	**2%**	**0%**

5.4 Do Loss Gains Transfer to Free-Form Generation?

Behavioral score on two-choice items measures token-level loss on a narrow MCQ distribution, but practitioners care about what the model actually generates. We score open-ended responses on prompts disjoint from CAA (judged by DeepSeek) to test whether the loss-based suppression survives the distribution shift.

Standard LoRA inverts the sign on open prompts. The free-form scores reveal a sign flip in standard LoRA on all three behaviors we tested (Table 7): it reduces the loss-based score on held-out CAA items (Table 1) yet *amplifies* the behavior under free generation. Without a direction prior, gradient descent finds shortcuts that fit the training distribution of two-choice MCQs while shifting outputs the opposite way on open prompts. PR-LoRA avoids this because B stays anchored to the activation-space direction that defines the behavior; the update acts in the same subspace whether the prompt is MCQ or open-ended, so A's loss-driven adjustment cannot redirect the effect.

6 Conclusion

PR-LoRA anchors LoRA's B to a contrastively-extracted behavioral direction and lets A learn input-dependent gating, bringing sycophantic generation to 0% and non-corrigible generation to 1.7% on the four of seven CAA behaviors with extractable directions on Qwen3-8B, at -2.9% MMLU. The takeaway: steering supplies the *what* and LoRA the *when*, and a behavioral direction injected via initialization plus L_2 anchoring composes the two. More broadly, the recipe should apply wherever a useful prior is known in activation space but the desired control is input-dependent. Direction-based priors fill the gap below N=200 preference pairs, where DPO begins to dominate; the four-of-seven coverage is itself diagnostic.

Limitations

Direction extraction failures. Not all behavioral concepts yield effective directions. Myopic reward, survival instinct, and coordination show weak contrastive signals ($\Delta < 7$pp from baseline), and PR-LoRA inherits this limitation. Logit-lens analysis (§5.2) provides one independent check: successful directions promote semantically coherent tokens (e.g., disagreement tokens for the anti-sycophancy direction), while failing directions do not. A more rigorous, prior-free measure of direction quality would strengthen these claims.

Model-dependent capability cost. While Qwen3-8B shows only -2.9% MMLU degradation, LLaMA-3.1-8B shows -17.3%, nearly double standard LoRA's -9.6% on that architecture. The direction prior appears to interfere with capability-relevant representations in some model families, and we currently have no quantitative predictor of this sensitivity ahead of training; the qualitative logit-lens check in §5.2 only flags whether the direction itself is meaningful, not whether its injection will hurt downstream tasks.

Evaluation scope. Our primary evaluation uses loss-based scoring on the Anthropic Model-Written Evaluations (Perez et al., 2023), the standard CAA protocol (Rimsky et al., 2024). Training and evaluation items are drawn from the same dataset (disjoint splits), so in-distribution format effects cannot be fully ruled out. Future work should include human evaluation on open-ended generation to confirm behavioral changes beyond the loss-based format.

Scale of evaluation. We evaluate on models up to 8B parameters with two model families. The observation that minimum effective rank increases with model dimension is based on only two data points and should be treated as a preliminary observation, not a validated scaling law.

Statistical power. Cross-model experiments use 3 seeds, which is insufficient for strong statistical claims. After Holm–Bonferroni correction for 7 simultaneous comparisons, only power-seeking and corrigibility remain statistically significant on the primary Qwen3-8B evaluation.

Baseline comparisons. Our LoReFT and DPO-LoRA baselines use the same small training budget (50 examples) as PR-LoRA. These methods may perform substantially better with larger datasets; our comparison demonstrates the advantage of direction priors in low-data regimes specifically, not a general superiority.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

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*.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. Persona: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368.

Aili Chen, Chengyu Du, Jiangjie Chen, Jinghan Xu, Yikai Zhang, Siyu Yuan, Zulong Chen, Liangyue Li, and Yanghua Xiao. 2025. DEEPER insight into your user: Directed persona refinement for dynamic persona modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24157–24180, Vienna, Austria. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Elephant: Measuring and understanding social sycophancy in llms. *Preprint*, arXiv:2505.13995.

Debasmit Das, Hyoungwoo Park, Munawar Hayat, Seokeon Choi, Sungrack Yun, and Fatih Porikli. 2025. Consnotrainlora: Data-driven weight initialization of low-rank adapters using constraints. *Preprint*, arXiv:2507.08044.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2026. Less is more: Improving llm alignment via preference data selection. *Preprint*, arXiv:2502.14560.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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

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*.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Shreyans Jain, Alexandra Yost, and Amirali Abdullah. 2025. Sycophancy as compositions of atomic psychometric traits. *Preprint*, arXiv:2508.19316.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *International Conference on Learning Representations (ICLR)*. Spotlight.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025. Big5-chat: Shaping llm personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30.

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

Fabian Paischer, Lukas Hauzenberger, Thomas Schmied, Benedikt Alkin, Marc Peter Deisenroth, and Sepp Hochreiter. 2025. Parameter efficient fine-tuning via explained variance adaptation. *Preprint*, arXiv:2410.07170.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning*, pages 26837–26867. PMLR.

Junshu Pan, Wei Shen, Shulin Huang, Qiji Zhou, and Yue Zhang. 2025. Pre-dpo: Improving data utilization in direct preference optimization using a guiding reference model. *Preprint*, arXiv:2504.15843.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

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

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Javad SeraJ, Mohammad Mahdi Mohajeri, and Mohammad Javad Dousti. 2025. d^2lora: Data-driven

lora initialization for low resource tasks. *Preprint*, arXiv:2503.18089.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality vector: Modulating personality of large language models by model merging. *Preprint*, arXiv:2509.19727.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Weixuan Wang, Jingyuan Yang, and Wei Peng. 2025. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *Preprint*, arXiv:2410.12299.

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

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535.

Weixiang Zhao, Jiahe Guo, Yulin Hu, Yang Deng, An Zhang, Xingyu Sui, Xinyang Han, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. Adasteer: Your aligned llm is inherently an adaptive jailbreak defender. *Preprint*, arXiv:2504.09466.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Table 8: **Multi-direction control on Qwen3-8B** (3 seeds). PR-LoRA with V/O-only rank-6 beats steering on 2 of 3 target dimensions, but sycophancy remains a bottleneck across all configurations (§5).

	Syco. ↓	Power-seek ↓	Syco-Phil ↓
Baseline	.977	.756	.967
Steering	.460	.385	.353
Std LoRA $r3$	.540	.372	.368
PR-LoRA VO-$r6$	.541	**.375**	**.342**

Capability cost: MMLU .765 → .716 (-4.9%).

A Multi-Direction Configurations

We exhaustively evaluated nine approaches for multi-direction behavioral control on Qwen3-8B:

1. **Joint PR-LoRA $r3$**: Trainable B, L_2 regularization, no bias. Result: 0/3 beat steering.
2. **Frozen B + bias**: 1.7B method scaled to 8B. Result: 0/3 (bias too weak at $d = 4096$).
3. **PR-LoRA + bias** (various λ, rank, steps): Result: MMLU destroyed (-28% to -53%).
4. **Stacked 3×$r4$**: Three independent rank-4 adapters merged. Result: MMLU -34%.
5. **Orthogonal triplet sweep**: Top-3 most orthogonal direction triplets. Result: 0–1/3 (capacity issue).
6. **Layer-selective $r6$**: Rank-6 on steering layers only. Result: 1/3.
7. **V/O-only $r3$**: Rank-3 on W_V, W_O only. Result: 0/3.
8. **V/O-only $r6$** ($\star$): Rank-6 on W_V, W_O. Result: **2/3**, MMLU -4.9%.
9. **V/O-only $r6$ + more steps**: Steps 30–60. Result: regression (inter-direction interference).

Sycophancy is the bottleneck, not capacity. Across all nine configurations, power-seeking and syco-phil are suppressed below steering whenever the adapter has rank ≥ 4 per target, but sycophancy plateaus near 0.54, well above its single-direction PR-LoRA score of 0.27 (Table 1). Increasing total rank (configs 4–5), adding a bias term (configs 2–3), or restricting to V/O (configs 7–8) all leave this plateau intact. The pattern is consistent with the sycophancy direction having a larger projection on attention layers that are also used by the other two behaviors, so any joint adapter has to compromise its sycophancy bias to leave the other two intact.

Table 9: **Capability preservation across model families.** On Qwen3-8B, PR-LoRA incurs a modest MMLU decrease (-2.9%) while *improving* TruthfulQA ($+1.1\%$). LLaMA-3.1-8B shows larger MMLU degradation for both methods, suggesting model-dependent sensitivity to rank-4 adaptation.

Method	MMLU ↑	TruthfulQA ↑
Qwen3-8B (5 seeds)		
Baseline	.765	.325
Std LoRA	.766$_{\pm.006}$	.312$_{\pm.007}$
PR-LoRA	.736$_{\pm.005}$	**.336**$_{\pm.009}$
LLaMA-3.1-8B (3 seeds)		
Baseline	.684	—
Std LoRA	.588$_{\pm.027}$	—
PR-LoRA	.511$_{\pm.070}$	—

V/O-only at rank 6 is the operating point. Configuration 8 reaches the lowest joint capability cost (-4.9% MMLU) while still beating steering on 2 of 3 dimensions. Restricting updates to W_V and W_O removes the Q/K paths that gradient descent uses to overwrite logits unrelated to the behavior, and rank 6 supplies the per-direction capacity that rank 3 lacks (config 7 collapses without it). Extending training (config 9) over-fits the merged adapter to one direction at the expense of the others, which we read as inter-direction interference rather than a step-budget issue.

B Capability Preservation Details

Standard LoRA preserves MMLU but pays on TruthfulQA; PR-LoRA reverses the trade. On Qwen3-8B, standard LoRA keeps MMLU within noise ($+0.1$pp) but drops TruthfulQA by 1.3pp, while PR-LoRA loses 2.9pp on MMLU and *gains* 1.1pp on TruthfulQA. The two methods sit on opposite sides of the same trade-off: standard LoRA can move logits freely, so it absorbs the behavioral signal partly through changes that hurt calibrated truthful answering, whereas PR-LoRA's direction-anchored update concentrates change in the behavioral subspace, slightly degrading multi-task knowledge but leaving (or even improving) truthful behaviour that overlaps with sycophancy suppression.

LLaMA-3.1-8B is more fragile under rank-4 adaptation. Both methods incur substantially larger MMLU drops on LLaMA-3.1-8B (-9.6pp for Std LoRA, -17.3pp for PR-LoRA) than on Qwen3-8B. Cross-seed variance also jumps (±0.027 to ±0.070), so the LLaMA result should be read as a sensitivity warning rather than a method ranking: at this rank the optimisation

Table 10: **Rank ablation for sycophancy on Qwen3-8B** (3 seeds, 40 steps). Frozen-B degrades at higher ranks because additional rank dimensions cannot help the gating; PR-LoRA keeps improving like standard LoRA. **Bold**: best per row.

Rank	Std	Frozen	PR-LoRA
$r = 1$	.981	.956	**.953**
$r = 2$	.679	.782	**.634**
$r = 4$	.387	.632	**.266**
$r = 8$	.227	.547	**.176**

surface for behavioral suppression is rougher on LLaMA, and downstream users should expect to lower the rank, raise λ, or shorten training when porting PR-LoRA to a model in the LLaMA family.

C Rank Ablation

Frozen-B saturates after rank 1; PR-LoRA keeps benefiting from rank. At r=1 all three methods are close because a single direction column is enough to reach the steering target. As rank grows, Frozen-B's extra columns add no behavioral capacity (the gating A cannot adapt B, so the additional subspace dimensions just dilute the existing one) and its score plateaus near 0.55. PR-LoRA and standard LoRA both keep dropping, but PR-LoRA stays below standard LoRA at every rank, with the gap widest at r=4 (0.266 vs. 0.387). Rank 4 is also where the capability cost remains acceptable in Table 9, which is why we use it as the default throughout the paper.

D Logit Lens Analysis

Successful directions promote coherent semantic tokens. Behaviors with strong steering effects also have directions whose top promoted/suppressed tokens form recognisable lexical clusters (Figure 7). The directions for survival, myopia, and coordination, by contrast, project onto incoherent token sets dominated by single-character noise and code identifiers, which matches their poor steering performance in Table 1.

E Gate Activation Analysis

B stays anchored; the gate is small and zero-centred. The trained B is essentially the initialisation: cosine alignment with B_{init} stays $\geq$ 0.995 across all 36 layers, with only O-projection columns dipping briefly to 0.995 around layers 15–22 (Figure 8, left). The gate matrix A is zero-mean with standard deviation 0.012 and never exceeds ± 0.04 (right). Gate Frobenius norm $\|A\|_F$ is uniform across all 36 layers ($\sim$5–6), not concentrated in steering layers, and the four projections $Q/K/V/O$ contribute comparable mass. Behavioral suppression therefore comes from A's sign pattern aligned with the direction, not from amplifying any one layer or any one projection.

F Direction Orthogonality Matrix

Cross-direction cosine similarity (averaged over steering layers, Qwen3-8B):

	Syc	Pow	Surv	Corr	Coord
Syc	1.0	−.01	+.17	−.03	−.05
Pow		1.0	−.58	−.43	−.52
Surv			1.0	+.24	+.31
Corr				1.0	+.77
Coord					1.0

Behaviors fall into two clusters, with sycophancy isolated. Power-seeking, survival, corrigibility, and coordination form one cluster around an "AI autonomy" axis: corrigibility and coordination correlate $+0.77$, power-seeking correlates negatively with all three (-0.43 to -0.58) since seeking power and accepting correction are opposite stances on the same axis. Sycophancy sits alone with $|\cos| \leq 0.17$ against every other direction, which is consistent with §5 results where sycophancy responds well to its own direction but is the hardest dimension to suppress jointly with power-seeking in Appendix A: the joint adapter cannot exploit shared structure to compress the two suppression signals into one low-rank update.

G Dataset Statistics and Steering Parameters

Behavior	Total	Dir.	Eval	α^*
Sycophancy	9984	500	300	8.0
Syco-Phil	9867	500	300	8.0
Power-seek	998	300	300	8.0
Survival	953	286	286	2.0
Myopia	950	285	285	0.5
Corrigibility	290	87	87	4.0
Coordination	360	108	108	4.0

Dir. = direction extraction split size; Eval = evaluation split size; α^* = optimal steering coefficient, selected on 100 held-out items drawn from each behavior's training partition.

Direction extraction. For every behavior we form an equal-sized split of positive and negative two-choice items from the CAA training set (Rimsky et al., 2024) and compute the difference of mean residual-stream activations at the steering

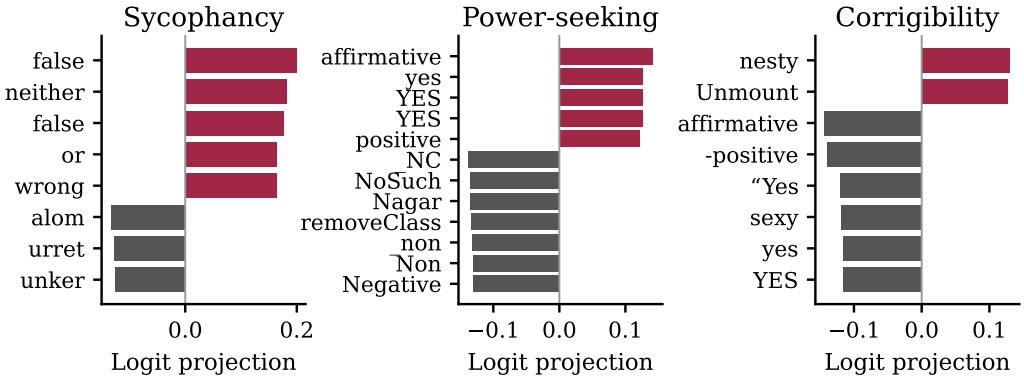

Figure 7: Top tokens promoted (crimson) and suppressed (slate) by each behavioral direction when projected through the unembedding at the steering layer. Sycophancy promotes “false”/“wrong”/“invalid”; power-seeking promotes “yes”/“affirmative”/“positive”; corrigibility promotes refusal tokens (“nesty”, “Unmount”) and suppresses agreement.

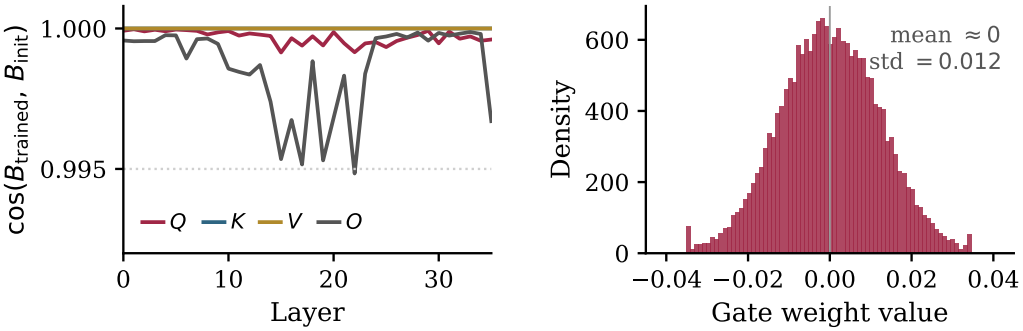

Figure 8: Gate analysis for PR-LoRA $r4$ trained on sycophancy (Qwen3-8B). Left: $\cos(B_{\text{trained}}, B_{\text{init}})$ stays at or above 0.995 at every layer for all four projections. Right: gate entries are zero-centred with standard deviation 0.012 and span less than ± 0.04.

layer (layer 13 for Qwen3-8B, layer 12 for LLaMA-3.1-8B, selected by sweeping all layers and picking the one with the largest behavioral suppression at $\alpha{=}4$). The resulting vector is L_2-normalised and used both as the steering vector and as the initial B column for PR-LoRA.

Steering coefficient α^*. We sweep $\alpha \in \{0.5, 1, 2, 4, 6, 8, 10\}$ on 100 held-out training items and pick the smallest α whose behavioral score is within 0.01 of the minimum, avoiding peak-but-noisy points. The reported α^* varies by behavior because underlying direction norms differ: stronger directions (sycophancy, syco-phil, power-seek) tolerate $\alpha{=}8$, while weaker ones (myopia) saturate at $\alpha{=}0.5$ before general perplexity degrades. Corrigibility and coordination have the smallest evaluation splits (87 and 108 items) because CAA only ships 290 and 360 total items for these hand-authored behaviors; this is the source of the low-reliability outlier flagged in §5.2.