LEARNING TO STEER, STEERING TO SEE: UNVEILING THE GEOMETRY OF RLVR IN LARGE LANGUAGE MODELS VIA TRAINABLE VECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has emerged as a key paradigm for enhancing the reasoning capabilities of large language models. However, the high dimensionality of parameter updates makes RL training dynamics difficult to analyze, obscuring the mechanisms underlying these gains. In this work, we study reinforcement learning with verifiable rewards (RLVR) and use vector steering as an analytical tool to identify a low-dimensional effective manifold in activation space that is closely associated with RL-induced performance gains. We further uncover two key geometric properties of this manifold. **(1) Effective Manifold Capacity:** The capacity required to reproduce RL-induced gains can be very small, but it is not infinitely compressible. When the capacity is compressed to an extremely low level, the intervention dimensionality and the ability to express input-dependent corrections become key constraints, and this requirement further varies with injection depth. **(2) Control Manifold Separation:** Effective control directions lie predominantly in the low-variance complement of the activation principal subspace. Within the same task and base model, the learned geometry remains largely consistent across training configurations; across tasks, geometric alignment correlates with capability transfer. We conduct experiments on 5 LLMs and 6 tasks with verifiable rewards, and the results support the above properties. Based on these findings, we propose **Alpha-Stabler**, a plug-and-play training framework comprising two modules: a **Predictor** that monitors principal-subspace intrusion to provide early warnings of training collapse, and a **Controller** that removes the principal-subspace component of activation gradients during backpropagation while preserving their components in the orthogonal complement of the principal subspace. Experiments show that Alpha-Stabler can stabilize training for 2,000 steps and consistently enhance the gains brought by reinforcement learning. This work advances the understanding of RL from an activation-manifold perspective and offers practical insights for more robust post-training. Our code is available at: https://anonymous.4open.science/r/alpha-stabler.

"Mille viae ducunt homines per saecula Romam."
— Alain de Lille

1 INTRODUCTION

Reinforcement learning (RL) has become central to post-training for large language models (LLMs), yielding substantial gains in reasoning performance (OpenAI, 2025; Xiao et al., 2026; Xu et al., 2026). Yet how RL produces these improvements internally remains incompletely understood. High-dimensional parameter updates, optimization noise, and sensitivity to training configurations make it difficult to distinguish behaviorally meaningful changes from incidental variation. Prior work has examined sampling efficiency (Yue et al., 2025), entropy dynamics (Cui et al., 2025b), scaling behavior (Tan et al., 2026), and update geometry (Cai et al., 2025; 2026; Zhu et al., 2025). A complementary question remains unanswered: what structure do RL-induced behavioral changes exhibit in activation space, and how does this structure relate to capability transfer and training stability?

We view RL post-training as an intervention problem in activation space. Concretely, we freeze the base model and learn input-invariant shared vectors at selected layers to distill its RL-tuned counterpart, thereby testing whether RL-induced gains can be recovered by low-dimensional activation interventions. This directly probes whether RL gains admit a low-dimensional description and when this description breaks down. We thereby identify a low-dimensional effective manifold in activation space tied to RL-induced gains, and characterize two geometric properties of this manifold: **Effective Manifold Capacity** and **Control Manifold Separation**.

Our first finding concerns the effective capacity of this manifold. With an RL-trained model as the teacher, a single input-invariant vector per controlled layer recovers over 85% of the gain achieved by full fine-tuning on most evaluated tasks, as shown in Fig. 1 (a) left. However, this compressibility has clear limits. The first arises from the teacher–student distribution gap: when the two distributions are close, a single-vector intervention suffices, whereas

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

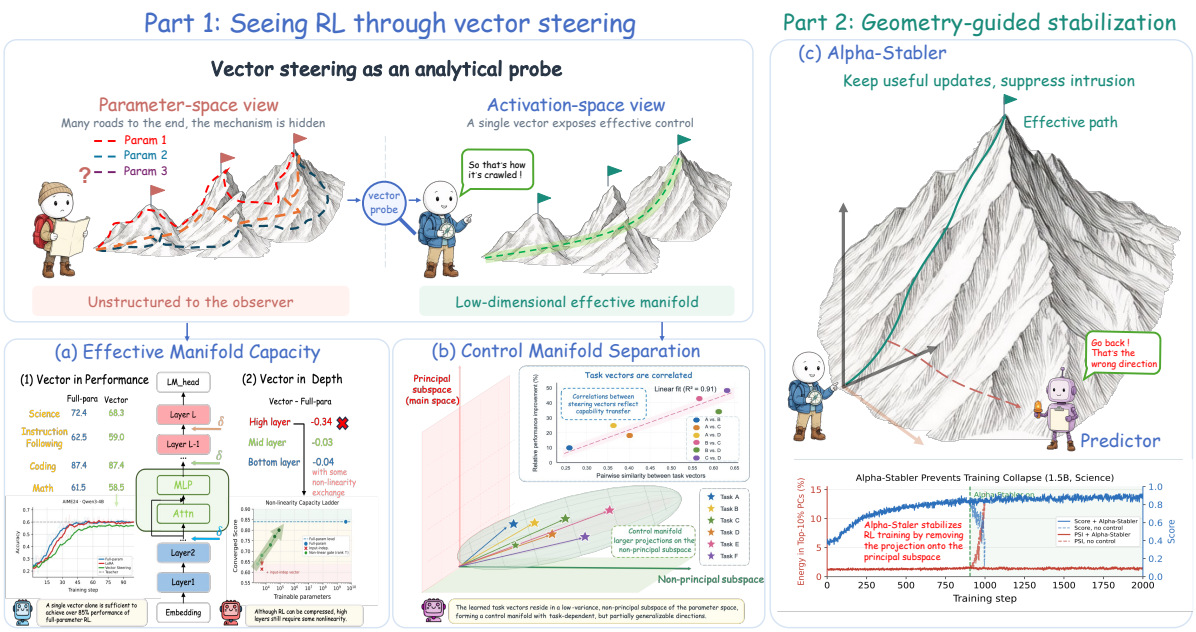

Figure 1: The single steering vector reveals that RLVR-induced changes form a low-dimensional, depth-sensitive control geometry, characterized by **(a)** effective manifold capacity and **(b)** control-manifold separation. Building on these findings, **(c)** Alpha-Stabler stabilizes training via principal-subspace intrusion monitoring and backward gradient projection.

a widening gap requires more flexible parameterizations. The second arises from insertion depth: fixed-vector interventions work well at low-to-middle layers but degrade near the output, as shown in Fig. 1(a) right. Through nonlinear analysis, we further find that this degradation does not stem from weak optimization signals at deeper layers, but from the expressiveness bottleneck of low-dimensional interventions at high layers. We summarize these findings as **Property 1 (Effective Manifold Capacity)**: the capacity required to reproduce RL-induced gains can be very small, but it is not infinitely compressible; this capacity depends on how well the intervention form matches the required correction.

We next examine the geometry of the effective manifold. By extending the single-vector probe to multiple mutually orthogonal vectors per layer, we find that although the standalone performance of later directions declines with extraction order, they can still independently support the target behavior, indicating that the leading direction is dominant but not uniquely effective. While RL's training parameters differ substantially across configurations in parameter space, they consistently converge, after compression, to a set of reproducible task-related directions. Across tasks, the alignment between direction vectors strongly reflects transfer gains, and these directions consistently concentrate in the low-variance complement of the activation principal subspace, as shown in Fig. 1 (b). We characterize this organization as **Property 2 (Control Manifold Separation)**: RL-induced control converges to a set of reproducible task-related directions that are largely separated from the variance-dominant representation subspace.

These two properties indicate that RL-induced behavioral changes admit a structured, low-dimensional description in activation space. During successful RL training, activation shifts remain predominantly in the low-variance complement of the principal subspace, whereas principal-subspace intrusion accompanies performance degradation and collapse. To this end, we propose **Alpha-Stabler**, a lightweight, plug-and-play framework for stabilizing RL training, as shown in Fig. 1 (c). It introduces no additional trainable parameters and requires no modification to the optimizer, reward function, or model architecture. Alpha-Stabler consists of a **Predictor** that monitors principal-subspace intrusion (PSI) against a fixed principal subspace from the frozen base model and warns before collapse, and a **Controller** that removes the principal-subspace component of activation gradients during backpropagation while preserving their orthogonal complements. Experiments show that, with no significant additional computation and zero additional trainable parameters, Alpha-Stabler achieves more stable reward growth over thousands of training steps than gradient clipping and other gradient projection methods.

In summary, this work identifies two geometric properties of RLVR-induced changes: **Effective Manifold Capacity** and **Control Manifold Separation**. These properties reveal that the seemingly complex dynamics of RL post-training are governed by a surprisingly simple structure: a low-dimensional manifold confined to the low-variance complement of the activation principal subspace. Building on these properties, we introduce **Alpha-Stabler**, a plug-and-play framework that operates with negligible overhead and provides a practical tool for stabilizing large-scale RL training.

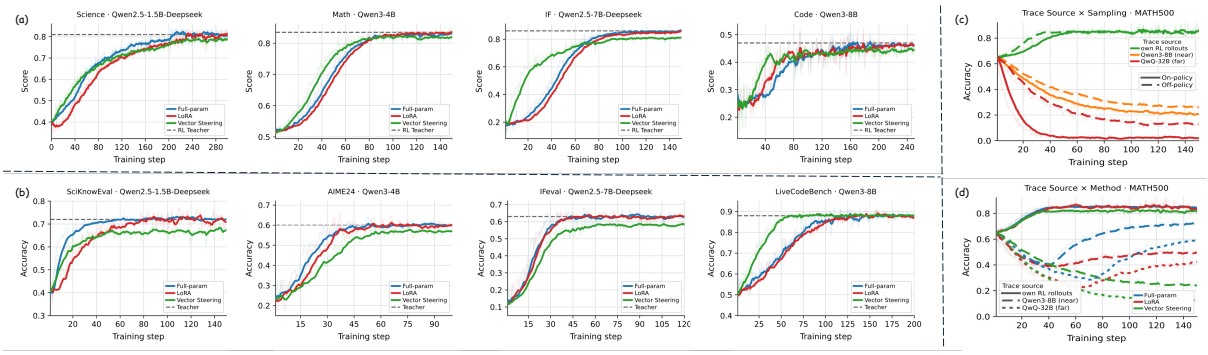

Figure 2: **(a)** On-policy distillation scores on the training set. **(b)** Off-policy distillation scores on the evaluation set. **(c)** Scores of vector steering under varying teacher types. **(d)** Scores of full-parameter training, LoRA, and vector steering under off-policy distillation at three distribution gaps.

2 Effective Manifold Capacity

2.1 Experimental Setting

Models and Datasets. We evaluate DeepSeek-R1-Distill-Qwen-1.5B/7B (Guo et al., 2025) and Qwen3-4B/8B/14B (Yang et al., 2025) on four task families: mathematical reasoning, scientific reasoning, code generation, and instruction following. For each task, we train a teacher with GRPO (Guo et al., 2025) or DAPO (Yu et al., 2025) and perform on- or off-policy distillation (Li et al., 2026b), comparing three update parameterizations: full-parameter training, LoRA (Hu et al., 2021), and vector steering. Details are provided in Appendix B.2.

Vector Steering. Consider a frozen base model with L decoder layers and hidden dimension d. For a set of controlled layers $\mathcal{S} \subseteq \{1, \ldots, L\}$, we add a trainable vector to each layer's output:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \boldsymbol{\delta}_\ell, \qquad \ell \in \mathcal{S}, \tag{1}$$

where $\mathbf{h}_\ell \in \mathbb{R}^d$ is the hidden state at a given token position, and $\boldsymbol{\delta}_\ell \in \mathbb{R}^d$ is a layer-specific vector shared across all inputs and token positions. Only $\{\boldsymbol{\delta}_\ell\}_{\ell \in \mathcal{S}}$ are optimized under the corresponding distillation objective. We then examine the compressibility of RL updates in Section 2.2 and further analyze its properties in Section 2.3.

2.2 Compressibility of RL Learning Signals

Compressibility across Distillation Regimes. We examine how much of the capability gain obtained by RL can be retained after distillation under restricted update parameterizations. As shown in Fig. 2 (a), under on-policy distillation, LoRA and full-parameter training perform comparably, while a single steering vector per controlled layer still recovers over 85% of the RL gain. To test whether this compressibility depends on student-generated trajectories (Yang et al., 2026), we also evaluate off-policy distillation with teacher-generated trajectories under the same update parameterizations. As shown in Fig. 2 (b), steering retains a comparably high recovery rate on evaluation datasets. Together, these results indicate that RL-induced gains can be highly compressed, and that fixed layer-wise activation offsets alone suffice to retain most of them.

Effect of Teacher Type and Update Capacity. These recovery rates, however, have limits. We fix the student and vary the teacher under both distillation regimes to examine the sensitivity of vector steering to the teacher–student distribution gap. As shown in Fig. 2 (c), vector steering effectively transfers the capabilities of RL-trained teachers, whereas steering performance degrades substantially with the tested non-RL teachers, with some settings collapsing as the gap widens. This indicates that teacher–student distribution matching is a key factor governing the effectiveness of fixed activation offsets.

A natural explanation is that the required corrections become more input-dependent as the gap widens, exceeding what an input-invariant fixed offset can express. If so, parameterizations with different expressive capacities should exhibit a consistent ordering. As shown in Fig. 2 (d), under off-policy distillation, the three parameterizations perform comparably when the gap is small, whereas a larger gap yields a clear ordering: full-parameter training performs best, followed by LoRA, and vector steering worst. This ordering matches their expressive capacities: full-parameter training allows model-wide adjustments, LoRA retains input-dependent corrections under low-rank constraints, and vector steering only applies fixed offsets. Therefore, a larger gap raises the input dependence required of the correction, and fixed additive corrections are the first to reach their capacity limit (Cai et al., 2026).

Implications for RL Compressibility. Taken together, these results show that, although vector steering cannot capture arbitrary teacher-induced corrections, it serves as a targeted probe for analyzing the mechanisms underlying

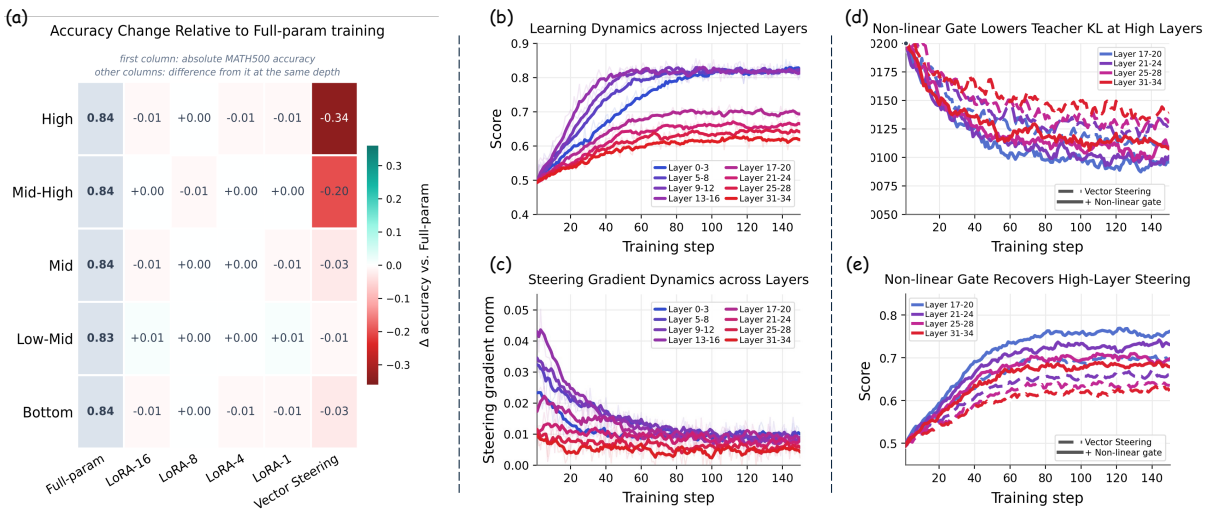

Figure 3: Layer-wise performance. **(a)** Converged performance of steering vectors across injection layers. **(b)** Task performance during training for steering vectors. **(c)** Steering gradient norms at different layers. **(d)** KL divergence with the teacher under fixed versus gated steering. **(e)** Converged performance of fixed versus gated injection.

RL-specific capability gains. Building on this, we next examine how RL-induced gains vary with injection depth and how effective the low-dimensional representation is across layers.

2.3 Layer-wise Compressibility of RL Updates

Injection Depth and Optimization Dynamics. Section 2.2 shows that RL-induced gains can be compressed into fixed activation offsets. This raises a complementary question: where should such a low-capacity intervention be applied? We therefore compare full-parameter updating, LoRA, and fixed-vector steering at representative network depths, restricting each intervention to the selected layer and freezing the remaining backbone. As shown in Fig. 3 (a), selected-layer full-parameter updating and LoRA remain effective across depths, whereas fixed-vector steering performs well at lower and intermediate layers but deteriorates near the output. This contrast indicates that low-capacity interventions depend on both the update parameterization and the injection depth.

We next examine the learning curves across injection sites. Fig. 3 (b) shows that shallow-layer steering gradually approaches intermediate-layer performance during training, whereas high-layer steering plateaus early, and extended training does not noticeably close the gap. To test whether this stagnation stems from weaker gradient signals, we inspect gradient norms. Fig. 3 (c) further shows that shallow and high-layer injections receive gradients of comparable magnitude yet yield substantially different final performance, indicating that gradient magnitude alone does not explain high-layer degradation. One explanation is that high-layer injections have fewer downstream nonlinear transformations, which may limit the ability of a fixed offset to produce context-dependent effects, whereas LoRA and selected-layer full-parameter updates can directly produce input-dependent activation corrections.

Input-Dependent Corrections and Capacity Controls. To test whether explicit token-dependent modulation alleviates high-layer degradation, we replace the fixed offset with a low-rank nonlinear correction:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \boldsymbol{\delta}_\ell(\mathbf{h}_\ell), \qquad \boldsymbol{\delta}_\ell(\mathbf{h}_\ell) = \mathbf{B}_\ell \sigma(\mathbf{A}_\ell \mathbf{h}_\ell)\,, \tag{2}$$

where $\mathbf{A}_\ell \in \mathbb{R}^{r\times d}$, $\mathbf{B}_\ell \in \mathbb{R}^{d\times r}$, $r \ll d$, and σ is an element-wise gating nonlinearity. Here $\mathbf{B}_\ell$ defines shared correction directions, while $\sigma(\mathbf{A}_\ell \mathbf{h}_\ell)$ assigns token-dependent coefficients. We zero-initialize $\mathbf{B}_\ell$ so that the intervention initially leaves the model unchanged. Fig. 3 (d,e) shows that Rank-1 gating reduces teacher KL divergence and improves converged performance at all four tested high-layer injection sites. At Rank-1, the correction still follows a single shared direction per layer, but its strength can vary across tokens. To separate this flexibility from parameter count, we compare against a parameter-matched input-independent control that replaces $\sigma(\mathbf{A}_\ell \mathbf{h}_\ell)$ with a constant coefficient. Fig. 4 (a) shows that the static control does not reproduce the improvement of Rank-1 gating, supporting the role of token-dependent modulation. Increasing the gating rank yields further gains toward the full-parameter reference at $r = 16$. Together, these results suggest that high-layer recovery depends on both the available correction directions and their input-dependent modulation.

Directional Structure and Effective Manifold Capacity. We further examine how gradient alignment and learned modulation vary across depth. At each layer, we compute the mean pairwise cosine similarity between per-token gradient contributions to the steering vector, $\mathrm{Sim}_\ell^{\mathrm{grad}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \cos(\mathbf{g}_{\ell,i}, \mathbf{g}_{\ell,j})$, where $\mathbf{g}_{\ell,i}$ denotes the normalized contribution of token i to the gradient of the loss with respect to the steering vector at layer ℓ, and

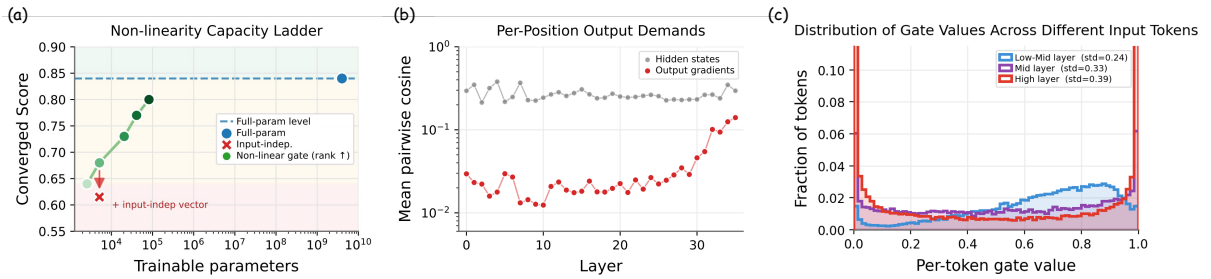

Figure 4: Diagnosing and mitigating high-layer degradation. **(a)** Converged performance of vector steering under different nonlinear parameters. **(b)** Layer-wise mean pairwise cosine similarity for hidden states and output gradients. **(c)** Distribution of gating activations at different layers.

$\mathcal{P}$ is the set of token pairs. We compute the same statistic for hidden states as a reference. Fig. 4 (b) shows that mean gradient similarity rises from about 0.019 at lower and intermediate layers to about 0.14 near the output. If directional alignment alone determined steering effectiveness, this trend would favor shared-direction interventions at higher layers. Yet fixed-vector steering degrades precisely there, showing that average directional alignment alone does not explain the depth-dependent performance gap. We further analyze the distribution of the Rank-1 gating coefficient $\sigma(\mathbf{A}_\ell \mathbf{h}_\ell)$ in Fig. 4 (c). At lower and intermediate layers, coefficients spread broadly over $[0, 1]$, consistent with graded modulation; at higher layers, they concentrate near 0 and 1, consistent with more selective use of the shared direction. Together with the gated-versus-static comparison in Fig. 4 (a), this indicates that explicit token-dependent modulation improves the effectiveness of high-layer interventions: when the intervention can adjust its strength according to each token's hidden state, the limited downstream nonlinear transformations at high layers no longer constitute a major bottleneck.

Summary. We summarize these findings as **Property 1 (Effective Manifold Capacity)**: reproducing RL-induced gains can require only a very low-dimensional intervention, but its effective capacity depends on how well the intervention form matches the required correction. Appendix D.4 provides a conditional account of fixed-vector recovery in terms of the balance between shared teacher-induced shifts and the context-dependent fluctuations around them.

3 Control Manifold Separation

Having established strong recovery with one steering vector per controlled layer, we now examine the diversity, reproducibility, and principal-subspace localization of effective steering directions.

3.1 Probing Residual Effective Directions with Multi-Vector Steering

Sequential orthogonal steering. To test whether the leading direction is uniquely effective, we sequentially learn K vectors $\mathcal{V}_\ell = \{\mathbf{v}_\ell^{(k)}\}_{k=0}^{K-1}$ at each controlled layer $\ell \in \mathcal{S}$, enforcing pairwise orthogonality against earlier vectors, $\langle \mathbf{v}_\ell^{(k)}, \mathbf{v}_\ell^{(j)} \rangle = 0$ for $j < k$. The first stage is trained to convergence and frozen; each subsequent stage initializes new vectors in the remaining orthogonal complement, optimizes them to convergence, and reprojects them after every optimizer step:

$$\mathbf{v}_\ell^{(k)} \leftarrow \mathbf{v}_\ell^{(k)} - \sum_{j=0}^{k-1} \frac{\left\langle \mathbf{v}_\ell^{(k)}, \mathbf{v}_\ell^{(j)} \right\rangle}{\left\| \mathbf{v}_\ell^{(j)} \right\|_2^2} \mathbf{v}_\ell^{(j)}. \tag{3}$$

At each stage, only the current stage's vectors are injected and evaluated, jointly defining a multi-layer intervention; earlier vectors remain frozen and serve solely to define the orthogonality constraints.

Residual effectiveness and model scale. As shown in Fig. 5 (a), the first direction achieves 67.5% standalone accuracy on SciKnowEval with DeepSeek-R1-Distill-Qwen-1.5B, while the eighth direction, $\mathbf{v}^{(7)}$, achieves 62.0%, only 5.5 percentage points lower despite being orthogonal to the preceding seven directions at every controlled layer. This modest reduction indicates that excluding previously learned directions does not exhaust the available steering solutions: the remaining orthogonal space still admits interventions with substantial standalone effectiveness. Training results across all four task domains are shown in Appendix Fig. 11.

We further examine how this phenomenon varies with model scale. As shown in Fig. 5 (b), the leading direction consistently performs best, while standalone accuracy generally declines with extraction order; the decline is sharper in smaller models, whereas larger models retain higher-order directions whose performance remains closer

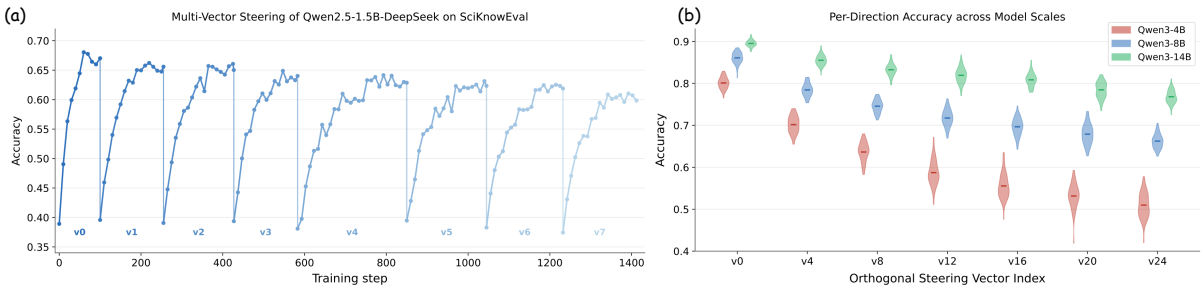

Figure 5: Multi-vector steering with sequential orthogonalization. **(a)** Training trajectories of sequentially learned orthogonal steering vectors on SciKnowEval with DeepSeek-R1-Distill-Qwen-1.5B. **(b)** Standalone accuracy of the extracted directions across model scales. Each direction is trained and evaluated independently, with preceding directions used only to define the orthogonality constraint.

to that of the leading direction. This indicates that model scale affects not only the performance of the leading direction, but also the extent to which higher-order directions are retained relative to it.

Summary. These results show that the leading steering direction is effective but not unique: multiple layer-wise orthogonal interventions can independently support the target behavior, although their standalone effectiveness generally declines with extraction order. We next examine whether comparable directions are recovered across teacher parameterizations and distillation regimes.

3.2 Task-Dependent Geometry and Principal-Subspace Separation

Consistency across training configurations. We next ask whether the extracted steering directions are reproducible across training configurations. We compare directions along two axes: teacher parameterizations (full-parameter fine-tuning vs. LoRA with different ranks) and distillation regimes (on- vs. off-policy). For each axis, we extract multi-vector directions using the same procedure and measure their layer-averaged cosine similarity, $\mathrm{sim}(\{\mathbf{v}_a^{(i)}\},\{\mathbf{v}_b^{(j)}\}) = \frac{1}{|\mathcal{S}|}\sum_{\ell\in\mathcal{S}}\cos(\mathbf{v}_{a,\ell}^{(i)},\mathbf{v}_{b,\ell}^{(j)})$, where $\mathcal{S}$ is the controlled layer set and i,j denote extraction indices. Fig. 6 (a) shows stronger same-index than cross-index alignment across the tested teacher parameterizations, while Fig. 6 (b) reports matching-index alignment between on- and off-policy directions. Together, these comparisons indicate a broadly consistent geometric structure across the evaluated configurations and distillation regimes within the same task.

Cross-topic alignment and capability transfer. Based on this, we further examine whether this geometric structure also reflects functional relationships across tasks. We partition Science into physics, chemistry, materials, and biology. Keeping the base model fixed, we train one RL teacher per topic and distill it using the same procedure to extract its multi-vector directions. For source topic s and target topic t, we define directional alignment as the sum of the first four same-index cosine similarities, averaged across controlled layers:

$$A(s,t) = \frac{1}{|\mathcal{S}|}\sum_{\ell\in\mathcal{S}}\sum_{i=0}^{3}\cos\left(\mathbf{v}_{s,\ell}^{(i)},\mathbf{v}_{t,\ell}^{(i)}\right).$$

We normalize the source-to-target transfer gain by the gain obtained from training directly on the target topic:

$$\Delta_{\mathrm{rel}}\,\mathrm{Acc}(s\to t) = \frac{\mathrm{Acc}(M_s,t)-\mathrm{Acc}(M_{\mathrm{base}},t)}{\mathrm{Acc}(M_t,t)-\mathrm{Acc}(M_{\mathrm{base}},t)}\times 100\%, \tag{4}$$

where M_s, M_t, and M_{base} denote the source-trained, target-trained, and base models, respectively. We evaluate six ordered source–target pairs among the four scientific topics, excluding self-pairs, and report the corresponding directional alignments and normalized transfer gains. As shown in Fig. 6 (c), stronger directional alignment is associated with larger normalized transfer gains across these pairs, with a linear fit yielding $R^2 = 0.91$. This indicates that the geometry of the learned steering directions is associated with capability transfer across the evaluated topic pairs, with stronger similarity corresponding to better transferability across tasks.

Effective steering directions avoid the principal subspace. We further examine where these directions are located in activation space. Let $U_{\ell,r}\in\mathbb{R}^{d\times r}$ contain the leading orthonormal eigenvectors of the frozen base model's token-level activation covariance at layer ℓ, with r corresponding to 10% of the hidden dimension, and define $P_\ell = U_{\ell,r}U_{\ell,r}^\top$. For a layer-wise direction $\mathbf{z}_\ell$, its principal-subspace energy fraction is

$$e_{\mathrm{PC}}(\mathbf{z}_\ell) = \frac{\|P_\ell\mathbf{z}_\ell\|_2^2}{\|\mathbf{z}_\ell\|_2^2+\epsilon}. \tag{5}$$

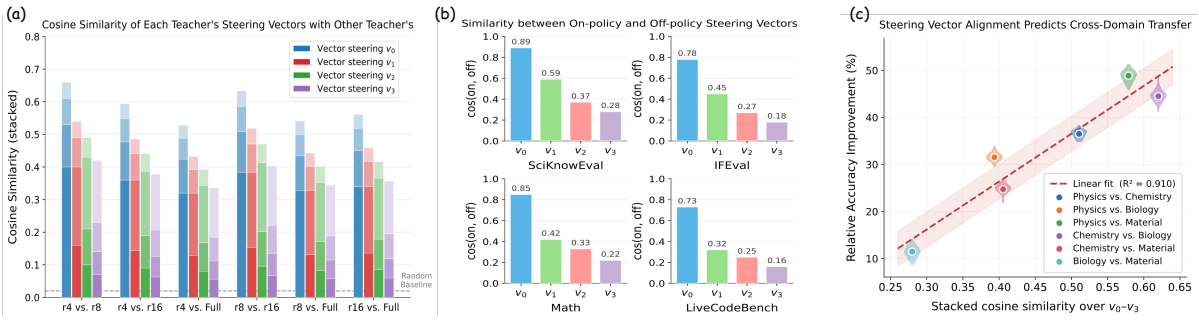

Figure 6: Geometric consistency and transfer of steering directions. **(a)** Layer-averaged cosine similarities between steering vectors under different teacher parameterizations, shown as stacked bars for each reference vector $\mathbf{v}_a^{(i)}$, with segments ordered from $\mathbf{v}^{(0)}$ to $\mathbf{v}^{(3)}$ indicating similarities with $\{\mathbf{v}_b^{(j)}\}_{j=0}^{3}$. **(b)** Layer-averaged cosine similarities between on- and off-policy steering vectors with matching extraction indices, per task. **(c)** Cross-topic directional alignment versus normalized source-to-target transfer gains, with a fitted regression line.

For multi-layer directions, we average e_{PC} across controlled layers. As shown in Fig. 7 (a), across four scientific topics, unconstrained steering places only about 1%–2% of its energy in the Top-10% principal subspace, well below the 10% isotropic reference, with this pattern persisting across the displayed directions $\mathbf{v}^{(0)}$ through $\mathbf{v}^{(8)}$. Low principal-subspace energy is thus a shared geometric property of the extracted interventions.

Restricting steering to the principal subspace does not recover the gain. To test the functional importance of the low-variance complement, we remove each steering vector's projection onto the bottom-70% activation-PC subspace after every optimizer step, restricting it to the span of the Top-30% components: $\mathbf{v}_\ell^{(k)} \leftarrow P_{\ell,m}\mathbf{v}_\ell^{(k)}$, where m is the number of retained components and $P_{\ell,m} = U_{\ell,m}U_{\ell,m}^\top$. As shown in Fig. 7 (b), the Top-10% energy fraction rises to about 15%, but accuracy does not improve meaningfully. The gain therefore does not come from alignment with the principal subspace.

Principal-subspace intrusion tracks training collapse. We next examine whether actual RL-induced activation shifts exhibit increased intrusion into the base model's Top-10% principal subspace as training becomes unstable. At step t, we run the current policy and the frozen base model on the same token sequences, defining the activation shift $\mathbf{\Delta}_{\ell,b,i}^{(t)} = \mathbf{h}_{\ell,b,i}^{(t)} - \mathbf{h}_{\ell,b,i}^{\mathrm{base}}(t)$ for sample b, position i, and layer ℓ. Here $\mathbf{h}_{\ell,b,i}^{\mathrm{base}}(t)$ is the frozen base model's activation on the token sequence used at update t; t indexes the sequence, not a change in base-model parameters. We stack all N valid token-level shifts row-wise into $X_\ell^{(t)} \in \mathbb{R}^{N\times d}$ to avoid cancellation before projection, and compute

$$\mathrm{PSI}_\ell^{(t)} = \frac{\|X_\ell^{(t)}U_{\ell,r}\|_F^2}{\|X_\ell^{(t)}\|_F^2 + \epsilon}, \tag{6}$$

where $U_{\ell,r}$ is the fixed Top-10% principal-subspace basis of the base model. As shown in Fig. 7(c), collapse-prone runs show a clear pattern: during stable training, PSI remains low, well below the 10% isotropic reference; as training proceeds, PSI departs from this range and becomes more volatile before any visible score degradation; as collapse progresses, PSI keeps rising while the score falls more sharply.

Summary. These results support **Property 2 (Control Manifold Separation)**: in the evaluated settings, effective steering directions lie mainly in the low-variance complement of the activation principal subspace, their ordered geometry stays broadly consistent across teacher parameterizations and distillation regimes, and cross-topic alignment correlates with transfer gains. Appendices D.5 and D.6 further discuss constrained vector fitting and the link between principal-subspace energy and output sensitivity.

4 Alpha-Stabler: Geometry-Guided RL Stabilization

Experiments show that effective steering directions lie in the low-variance complement, while principal-subspace intrusion accompanies training collapse. We therefore propose Alpha-Stabler to monitor whether the RL-induced activation shift departs from this regime. It consists of a Predictor, which detects sustained increases in principal-subspace intrusion (PSI), and a Controller, which removes the principal-subspace component of activation gradients during backpropagation while preserving each incoming gradient's component in the orthogonal complement of the principal subspace.

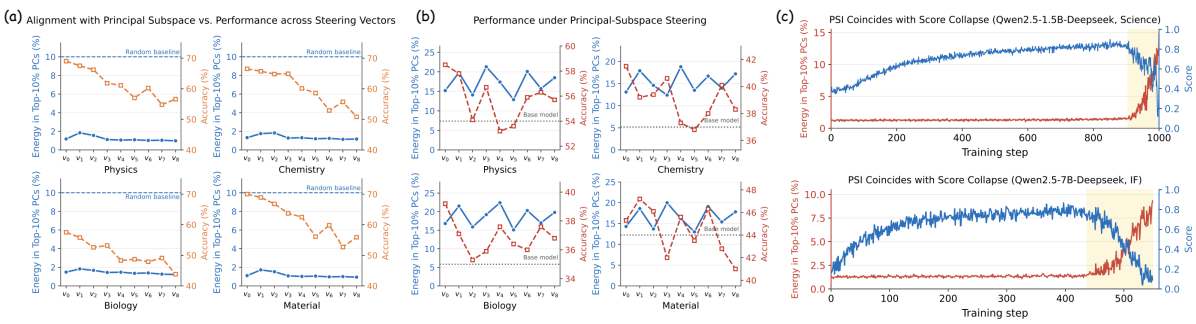

Figure 7: Steering geometry and principal-subspace intrusion. Energy fractions in **(a)** and **(b)** use the Top-10% principal subspace. **(a)** Standalone accuracy and principal-subspace energy fractions of sequentially extracted directions across four scientific topics. **(b)** The same quantities when steering vectors are restricted to the Top-30% activation-PC subspace. **(c)** Task score and PSI during RL training, comparing stable and collapse-prone runs.

4.1 Online Monitoring and Control

At parameter-update step t, we evaluate the actor and the frozen initial base model on identical token sequences. We stack their valid token-level differences, $\Delta_{\ell,b,i}^{(t)} = \mathbf{h}_{\ell,b,i}^{(t)} - \mathbf{h}_{\ell,b,i}^{\text{base}}(t)$, into $X_\ell^{(t)} \in \mathbb{R}^{N\times d}$. During the first $T_{\text{warm}} = 50$ updates, actor-generated rollouts are evaluated by the frozen base model to estimate the base model's token-level activation covariance at each monitored layer. At the end of warm-up, we extract its leading orthonormal eigenvectors $U_{\ell,r}$, with $r = \lceil 0.10d \rceil$, and keep these bases fixed thereafter. Using the layer-wise PSI defined in Eqn. 6, we additionally aggregate across monitored layers:

$$\text{PSI}^{(t)} = \frac{\sum_{\ell\in\mathcal{S}} \|X_\ell^{(t)} U_{\ell,r}\|_F^2}{\sum_{\ell\in\mathcal{S}} \|X_\ell^{(t)}\|_F^2 + \epsilon}. \tag{7}$$

Layer-wise PSI determines where control is activated; the aggregate summarizes intrusion across monitored layers according to their shift energies. Both sum token-level energies before normalization, avoiding cancellation between differently oriented shifts. PSI is computed from detached activations, so it does not affect gradients, while the actor's computation graph is retained for backpropagation.

Predictor: detecting sustained intrusion. During an initially stable warm-up, the actor trains without projection, and valid activation-shift matrices are retained every M updates. Once the principal bases are fixed, we compute PSI for these retained shifts and set fixed warning and lower release thresholds from layer-wise medians and scaled median absolute deviations. Subsequent valid observations update an exponential moving average initialized at the warm-up median. Control activates after p consecutive valid checks with the smoothed PSI above the warning threshold, and releases when it falls below the lower threshold. Algorithm 1 specifies the full procedure.

Controller: projecting activation gradients. The Controller modifies backward gradients without changing forward activations. Let $G_\ell^{(t)} \in \mathbb{R}^{N\times d}$ denote the incoming gradient at layer ℓ, including any downstream control effects. With the Predictor's flag $a_\ell^{(t)} \in \{0, 1\}$ fixed during backpropagation and the optimizer step, we apply

$$\tilde{G}_\ell^{(t)} = G_\ell^{(t)} - a_\ell^{(t)} \left(G_\ell^{(t)} U_{\ell,r}\right) U_{\ell,r}^\top. \tag{8}$$

When active, this removes the component of each token's gradient in the principal subspace while preserving the component orthogonal to that subspace. It integrates directly into standard RL training: rollout generation, likelihood ratios, rewards, and the GRPO or DAPO loss remain unchanged, and the optimizer updates parameters using only the modified backward signal. Warm-up requires estimating reference statistics and temporarily storing calibration shifts; after calibration, reference activations are needed only at monitoring checks and can be reused from a compatible pass. Each active hook performs two low-rank matrix multiplications, and inference uses the trained actor alone, so the additional computation is minimal. Algorithm 1 summarizes the full training loop.

4.2 Experimental Evaluation

Training stabilization. Fig. 8 (a) and (b) compares training trajectories with and without Alpha-Stabler.. In uncontrolled runs, a pronounced increase in PSI coincides with sharp score deterioration, indicating that performance loss accompanies a departure from the earlier activation geometry. Under control, PSI remains near its early-training range while task scores continue to improve, yielding sustained learning alongside limited principal-subspace intrusion rather than reduced PSI at the expense of task performance.

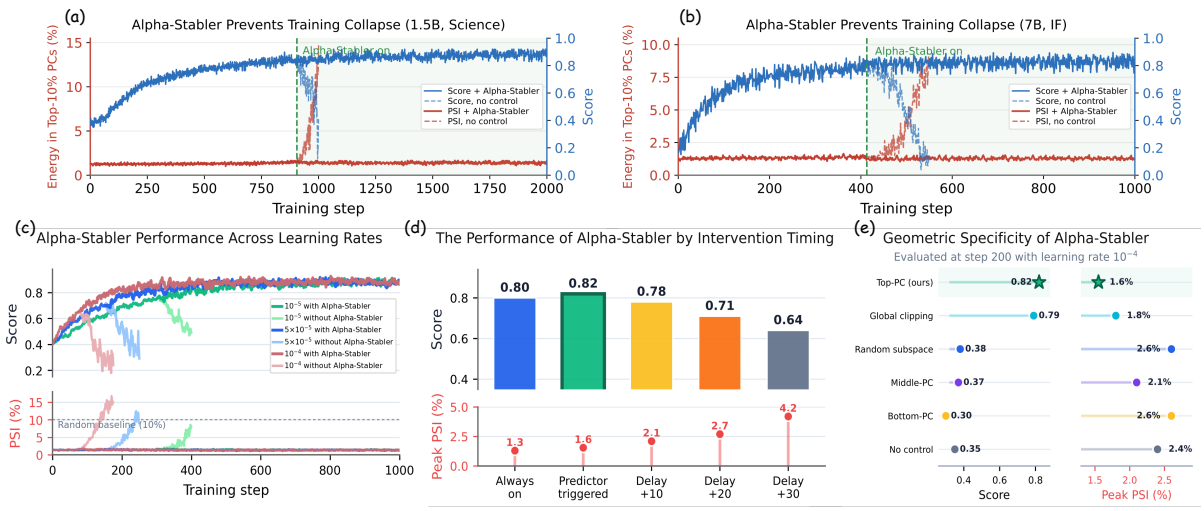

Figure 8: **(a, b)** Training dynamics with and without Alpha-Stabler. **(c)** Training dynamics across learning rates. **(d)** Task score and peak PSI under different intervention timings. **(e)** Performance under control of different subspaces.

Ablation studies. We then conduct ablation studies of Alpha-Stabler along three dimensions: learning rate, intervention timing, and controlled subspace. **First**, we find that stabilization holds across learning rates, as shown in Fig. 8 (c): at each tested learning rate, uncontrolled training eventually collapses, whereas controlled training keeps peak PSI at a low single-digit percentage and reaches comparable final scores across settings, indicating robustness to the choice of learning rate. **Second**, triggered control is preferable to delayed control, as shown in Fig. 8 (d). Triggered control matches always-on control in both score and peak PSI, indicating that continuous intervention is unnecessary, whereas delaying intervention progressively degrades the score and raises peak PSI. **Third**, Top-PC projection outperforms alternative subspaces, as shown in Fig. 8 (e). Among the compared gradient-control methods, removal of Top-PC components achieves the highest score, while random-subspace, Middle-PC, and Bottom-PC control fail to match it and remain close to the uncontrolled baseline; Top-PC also compares favorably with global gradient clipping. This indicates that the improvement does not arise from an arbitrary gradient constraint, but from removing gradient components in a specific subspace, namely the principal subspace. These results support geometrically targeted control rather than indiscriminate gradient suppression. Appendix D.7 shows that gradient projection minimally removes the principal-subspace component of the incoming gradient and bounds principal-subspace activation leakage for a local, single-site projected step. Fig. 12 further shows nearly identical score trajectories with and without Alpha-Stabler under the same measured wall-clock time, indicating almost no additional time overhead.

5 Limitations and Future Work

Despite revealing two geometric properties of RLVR's effect on LLMs, our study has several limitations. First, the theoretical framework rests on empirically motivated assumptions not yet derived from first principles. Second, our analysis and experiments are restricted to RLVR on tasks such as mathematical reasoning and code generation; we do not validate the proposed geometry on RLHF, multi-turn agentic decision-making, or open-ended generation, where whether the same properties hold remains open. Future work can relax these assumptions via neuron attribution and causal tracing, extend the framework to RLHF and agentic settings, and exploit off-principal localization as a broader design principle, e.g., by adding complement-space regularizers or using PSI as an adaptive signal for learning rate and intervention strength, enabling low-rank control without layer monitoring.

6 Conclusion

In this work, we identify two geometric properties of RLVR's effect on LLMs. Effective Manifold Capacity characterizes how a very low-dimensional intervention can reproduce RLVR-induced gains, with the required capacity depending on how well the intervention form matches the required correction. Control Manifold Separation describes the concentration of reproducible, task-related control directions in the low-variance complement of the activation principal subspace. Based on these properties, Alpha-Stabler monitors principal-subspace intrusion as an early-warning signal and projects away the intrusive component of activation gradients during backpropagation, stabilizing training with almost no additional computational overhead. Overall, this work recasts RLVR from black-box parameter optimization into an interpretable geometric process, opening new avenues for theoretically grounded and practically efficient LLM post-training.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and formatting, with the aim of improving clarity, readability, and presentation. All AI-assisted revisions were critically reviewed and revised by the authors to ensure technical accuracy, consistency, and fidelity to the underlying research. The authors take full responsibility for the final manuscript, including its methodology, theoretical arguments, experimental results, and conclusions.

REPRODUCIBILITY STATEMENT

We document the experimental procedures and theoretical assumptions underlying our results. Appendix B specifies the training objectives, model configurations, datasets, task-specific verifiers, evaluation protocols, prompts, optimization hyperparameters, and hardware, together with example commands for vector-steering experiments. Appendix C provides additional analyses of the learned steering vectors. Appendix D states the assumptions and proofs supporting our geometric analysis and details the principal-subspace diagnostic and gradient-projection mechanism of Alpha-Stabler. Algorithm 1 specifies its warm-up calibration, monitoring, and control procedures. The source code is provided in the supplementary materials.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37: 136037–136083, 2024.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guangzhong Sun, Guiquan Liu, and Junfeng Fang. On predictability of reinforcement learning dynamics for large language models. *arXiv preprint arXiv:2510.00553*, 2025.

Yuchen Cai, Ding Cao, Liang Lin, Chunxi Luo, Xin Xu, Kai Yang, Weijie Liu, Saiyong Yang, Tianxiang Zhao, Guangzhong Sun, et al. Learning to foresee: Unveiling the unlocking efficiency of on-policy distillation. *arXiv preprint arXiv:2605.11739*, 2026.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025b. URL https://arxiv.org/abs/2505.22617.

Yifeng Ding, Hung Le, Songyang Han, Kangrui Ruan, Zhenghui Jin, Varun Kumar, Zijian Wang, and Anoop Deoras. Empowering multi-turn tool-integrated agentic reasoning with group turn policy optimization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42409–42423, 2026.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *Advances in Neural Information Processing Systems*, 38:36256–36282, 2026.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, 2022. URL https://arxiv.org/abs/2203.14680.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. URL https://arxiv.org/abs/2504.11456.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Kun Li, Zenan Xu, Junan Li, Zengrui Jin, Jinghao Deng, Zexuan Qiu, and Bo Zhou. Discovery and reinforcement of tool-integrated reasoning chains via rollout trees. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14265–14283, 2026a.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

Alex Gu Wen-Ding Li Fanjia Yan Tianjun Zhang Sida Wang Armando Solar-Lezama Koushik Sen Ion Stoica Naman Jain, King Han. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint*, 2024.

OpenAI. Introducing gpt-oss. https://openai.com/zh-Hans-CN/index/introducing-gpt-oss/, 2025.

Simon Ostermann, Daniil Gurgurov, Tanja Baeumel, Michael A Hedderich, Sebastian Lapuschkin, Wojciech Samek, and Vera Schmitt. From weights to activations: Is steering the next frontier of adaptation? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29854–29879, 2026.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, august 2026. URL https://qwen.ai/blog?id=qwen3.8.

Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and Dongrui Liu. Rethinking generalization in reasoning sft: A conditional analysis on optimization, data, and model capability. *arXiv preprint arXiv:2604.06628*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, Zhongzhi Li, Zaibin Zhang, Guibin Zhang, Chen Zhang, Zhenfei Yin, Philip Torr, and Lei Bai. Scaling behaviors of llm reinforcement learning post-training: An empirical study in mathematical reasoning, 2026. URL https://arxiv.org/abs/2509.25300.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

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

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Y. Yang, Y. He, J. Liu, and Z. Jin. Making complex reasoning student-friendly: A hybrid llm-to-slm distillation framework. In *1st Workshop on Scaling Post-training for LLMs at the Fourteenth International Conference on Learning Representations (ICLR)*, April 2026. URL https://openreview.net/forum?id=6jP5PDOmqN.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Z.ai. Glm-5.3: Frontier coding with emergent cyber capabilities, august 2026. URL https://z.ai/blog/glm-5.3.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024. URL https://arxiv.org/abs/2405.01470.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025.

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, et al. The path not taken: Rlvr provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

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

A Related Work

Reinforcement Learning for Large Language Models. Prior to reasoning-capable models like OpenAI's o1, RL was primarily used in RLHF for instruction-following and human preference alignment (Ouyang et al., 2022). Recently, Reinforcement Learning with Verifiable Rewards (RLVR) has proven effective for reasoning in verifiable domains such as mathematics and programming (Lambert et al., 2025). OpenAI's o1 inspired a wave of open reasoning models, including DeepSeek-V4 (Xu et al., 2026), Kimi-K3 (Team et al., 2026), Qwen3.8 (Qwen Team, 2026), and GLM-5.3 (Z.ai, 2026). Beyond single-turn reasoning, RL is increasingly applied to agentic settings involving planning, tool invocation (Li et al., 2026a), and multi-turn environment interaction (Ding et al., 2026), with notable efforts on search-augmented reasoning (Jin et al., 2025), multi-turn credit assignment (Zhou et al., 2025), and asynchronous large-scale agentic RL (Fu et al., 2026).

Emergent Behaviors of On-Policy Training. Recent studies have investigated intrinsic properties of on-policy RL training. Yue et al. (2025) find that RL primarily improves pass@1 sampling efficiency without expanding the model's reasoning frontier. Cui et al. (2025b) identify early-stage entropy collapse that prematurely degrades exploration. Tan et al. (2026) reveal a power-law relationship between test loss, computation, and data volume, showing larger models exhibit superior learning efficiency. From a parameter-dynamics perspective, Cai et al. (2025) uncover Rank-1 dominance and Rank-1 linear dynamics, enabling 2.5× acceleration via AlphaRL. Cai et al. (2026) attribute on-policy distillation's efficiency to modular redundancy suppression and early directional stabilization, proposing EffOPD with 3× speedup (Song & Zheng, 2026).

Steering Vectors and Representation Engineering. While most adaptation methods rely on parameter updates (full fine-tuning or LoRA), an alternative line of work intervenes directly on internal activations via vector steering. Representation Engineering (RepE) provides a unifying framework, treating population-level representations as the analytical unit for monitoring and manipulating cognitive phenomena (Zou et al., 2023). High-level behaviors often exhibit low-dimensional structure: refusal is mediated by a single direction (Arditi et al., 2024) or multi-dimensional concept cones (Wollschläger et al., 2025). Task vectors (Ilharco et al., 2022) and function vectors (Todd et al., 2024) enable behavioral editing through vector arithmetic, and steering is increasingly viewed as a paradigm complementary to fine-tuning and prompting (Ostermann et al., 2026).

B PRELIMINARIES AND EXPERIMENTAL SETUP

B.1 PRELIMINARIES

We consider three post-training paradigms: Reinforcement Learning (RL), On-Policy Distillation (OPD), and Off-Policy Distillation. Let π_θ denote the policy or student model, π^* the fixed teacher model, and π_{ref} the fixed reference model used for regularization. The teacher provides distillation targets, whereas the reference model anchors RL regularization; these roles are distinct.

Reinforcement Learning (RL). The RL objective is:

$$\max_\theta J_{\text{RL}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\right] - \beta D_{\text{KL}}\left(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\right)\right], \tag{9}$$

where $y = (y_1, \ldots, y_T)$ is sampled from the current policy, $r(x, y)$ is the reward assigned to response y for query x, and β controls the strength of KL regularization.

In Reinforcement Learning from Verifiable Rewards (RLVR), $r(x, y)$ is determined by verifiable outcomes, such as answer correctness or unit-test execution, without requiring a learned reward model.

Separating the task-reward term from reference-policy regularization gives the policy-gradient form:

$$\begin{aligned}\nabla_\theta J_{\text{RL}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\left[\sum_{t=1}^{T} A_t^r \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\right] \\ - \beta \nabla_\theta \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}\left(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\right)\right],\end{aligned} \tag{10}$$

where A_t^r denotes the token-level advantage associated with the task reward. In RLVR, this reward is typically observed only after the complete response has been generated. Equation 10 describes the underlying regularized objective; practical GRPO and DAPO updates use algorithm-specific advantage estimates and clipped surrogate objectives.

On-Policy Distillation (OPD). OPD samples trajectories from the student policy and minimizes the reverse KL divergence from the student to the teacher:

$$\begin{aligned}\min_\theta J_{\text{OPD}}(\theta) &= \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}\left(\pi_\theta(\cdot \mid x) \,\|\, \pi^*(\cdot \mid x)\right)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\left[\sum_{t=1}^{T}\left(\log \pi_\theta(y_t \mid x, y_{<t}) - \log \pi^*(y_t \mid x, y_{<t})\right)\right].\end{aligned} \tag{11}$$

The corresponding sequence-level gradient can be written as:

$$\begin{aligned}\nabla_\theta J_{\text{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\Bigg[\sum_{t=1}^{T}\sum_{t'=t}^{T}\left(\log \pi_\theta(y_{t'} \mid x, y_{<t'}) - \log \pi^*(y_{t'} \mid x, y_{<t'})\right) \\ \cdot \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\Bigg].\end{aligned} \tag{12}$$

For optimization, we use an immediate token-level approximation that retains only the contribution at $t' = t$:

$$\begin{aligned}\nabla_\theta J_{\text{OPD}}(\theta) \approx \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\Bigg[\sum_{t=1}^{T}\left(\log \pi_\theta(y_t \mid x, y_{<t}) - \log \pi^*(y_t \mid x, y_{<t})\right) \\ \cdot \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\Bigg].\end{aligned} \tag{13}$$

This approximation provides token-level supervision along student-generated trajectories, but omits the future-token contributions in Equation 12. When implemented as an advantage-weighted surrogate, the log-probability difference is treated as a stop-gradient weight, with gradients taken through the student log-probability that it multiplies.

Off-Policy Distillation (Off-PD). Off-policy distillation instead samples trajectories from the teacher policy. Its objective is the forward KL divergence from the teacher to the student:

$$\min_\theta J_{\text{Off-PD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}\left(\pi^*(\cdot \mid x) \,\|\, \pi_\theta(\cdot \mid x)\right)\right]$$

$$= \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi^*(\cdot|x)} \left[\sum_{t=1}^{T} \left(\log \pi^*(y_t \mid x, y_{<t}) - \log \pi_\theta(y_t \mid x, y_{<t}) \right) \right]. \tag{14}$$

Since the teacher-dependent term is constant with respect to θ, this objective is equivalent to minimizing the negative log-likelihood of teacher-generated trajectories:

$$\mathcal{L}_{\text{Off-PD}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi^*(\cdot|x)} \left[\sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t}) \right]. \tag{15}$$

The corresponding gradient is:

$$\nabla_\theta \mathcal{L}_{\text{Off-PD}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi^*(\cdot|x)} \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}) \right]. \tag{16}$$

Thus, OPD and off-policy distillation differ in both trajectory sampling and the objective being optimized. OPD uses student-generated trajectories with a reverse-KL objective, whereas off-policy distillation uses teacher-generated trajectories with a forward-KL objective, implemented as sequence negative log-likelihood. The latter trains the student on teacher-generated prefixes, while the former evaluates the teacher on prefixes generated by the student.

B.2 Experimental Setup

Overview. Each experiment in this paper is defined by a training paradigm and an update configuration. The three paradigms are on-policy RL with verifiable rewards (Eqn. 9), on-policy distillation (Eqn. 11), and off-policy distillation (Eqn. 15), as formalized in Section B.1. For RL, we consider full fine-tuning and LoRA, whereas for both on-policy and off-policy distillation, we additionally evaluate vector steering. These configurations differ in which parameters receive gradients and, consequently, in the dimensionality and expressiveness of the resulting updates. Within each matched comparison, we hold the base model, prompt set, teacher model when applicable, sampling recipe, and evaluation protocol fixed across the applicable update configurations. Parameterization-specific optimization settings are reported in Table 2, and execution differences are described below. In on-policy experiments, matching the sampling recipe does not imply identical trajectories, since the samples depend on the current policy. This setup allows us to compare update parameterizations under matched modeling and evaluation protocols while making their optimization differences explicit.

Models. To ensure the generality of our findings, we conduct experiments across model scales ranging from 1.5B to 14B parameters, spanning two model families: Qwen2.5/DeepSeek-R1-Distill (1.5B and 7B) and Qwen3 (4B, 8B, and 14B), and two RL algorithms (GRPO and DAPO). All RL models are trained locally using the `verl` framework (Sheng et al., 2025). For all distillation students, the default teacher is the RL-tuned version of the student's own base model, ensuring that the teacher–student gap corresponds exactly to the RL update itself.

Datasets and Evaluation. The training data spans four topics, each with its own verifier and held-out evaluation set (Table 1). Rewards are always *verifiable* in the sense of §B.1: rule-based answer matching for mathematical and scientific reasoning, unit-test execution for code, and programmatic constraint checking for instruction following. For the off-policy distillation experiments in Fig. 2 (d), we additionally compare two sources of supervision: standard SFT on Math-CoT-44K (Ren et al., 2026), which contains approximately 44K mathematical queries and 32 responses sampled from Qwen3-32B for each query, yielding more than 1.4M reasoning traces; and off-policy distillation on OpenMathReasoning (Moshkov et al., 2025), which contains 540K mathematical problems and 3.2M long chain-of-thought solutions generated by DeepSeek-R1 and QwQ-32B.

Table 1: Training and evaluation datasets and reward functions for each domain.

Domain	Training Set	Evaluation Set	Reward
Math	DeepMath-103K (He et al., 2025)	AIME 2024 (Ye et al., 2025) MATH500 (Lightman et al., 2023)	Answer matching
Science	SciKnowEval-Training (Feng et al., 2024)	SciKnowEval-Eval (Feng et al., 2024)	Answer matching
Code	Eurus (Cui et al., 2025a)	LiveCodeBench-v5 (Naman Jain, 2024)	Unit tests
Instruction following	IFEval-Training (Zhao et al., 2024)	IFEval-Eval (Zhao et al., 2024)	Constraint checking

Instruction-following split. For instruction following, we partition the dataset at the prompt level into an 80% training split and a 20% held-out evaluation split, denoted as IFEval-Training and IFEval-Eval, respectively. These names refer to the partitions used in this study. All responses associated with the same prompt remain in the same partition, and evaluation prompts are excluded from training. The split is kept fixed across all compared configurations.

Evaluation protocol. We report `mean@4` on the held-out evaluation sets. For each evaluation prompt, we independently sample four responses using the same decoding configuration. Each response is scored using the corresponding task-specific evaluation criterion. The reported result is the arithmetic mean of the four single-sample evaluation scores computed on the same held-out set:

$$\text{mean@4} = \frac{1}{4}\sum_{j=1}^{4} \text{Score}\left(\{(x_i, y_i^{(j)})\}_{i=1}^{N}\right), \tag{17}$$

where N is the number of evaluation prompts, $y_i^{(j)}$ is the j-th sampled response to prompt x_i, and Score is the task-specific evaluation metric. For mean per-response correctness, this is equivalently the average over all $4N$ scored responses. We do not select the best of the four responses or apply majority voting; `mean@4` is therefore distinct from `pass@4`.

Update configurations. We compare the following trainable parameterizations under the matched modeling and evaluation protocols described above, using the configuration-specific learning rates in Table 2.

- **Full fine-tuning** updates all model parameters by default and serves as the unrestricted reference configuration. In principle, it can express a substantially broader class of parameter updates than the other configurations.
- **LoRA** (Hu et al., 2021) injects low-rank adapters on `all-linear` target modules while keeping the backbone frozen. This restricts the update to a low-rank form, though each input can still produce a different correction (i.e., it remains context-dependent). The rank varies across experiments; Table 2 reports the default rank-8 configuration.
- **Vector steering** freezes the entire backbone and trains context-invariant bias vectors added to the residual stream. For a controlled layer set $\mathcal{S}$, the single-vector intervention is

$$\mathbf{h}_\ell \;\mapsto\; \mathbf{h}_\ell + \mathbf{v}_\ell, \qquad \mathbf{v}_\ell \in \mathbb{R}^d, \quad \ell \in \mathcal{S}, \tag{18}$$

  with $|\mathcal{S}|d$ trainable parameters. For the default single-vector configuration on Qwen3-4B ($d = 2560$) with $\mathcal{S} = \{5, \ldots, 20\}$ and $|\mathcal{S}| = 16$, this amounts to 40,960 scalars, approximately 4.1×10^4, or roughly 1.0×10^{-5} of the backbone. Crucially, $\mathbf{v}_\ell$ does not vary with the input: the *same* vector is added regardless of context or decoding position.

Unless otherwise stated, vector steering uses the controlled layer set $\mathcal{S} = \{5, \ldots, 20\}$ and a single vector per controlled layer ($K = 1$). In the layer-wise experiments of Section 2.3, all three configurations are restricted to the selected layer: full-parameter updating trains only that layer, LoRA adapts its linear modules, and vector steering adds an offset to its output. The remainder of the backbone is frozen. These selected-layer experiments are distinct from the default model-wide full-fine-tuning and LoRA configurations.

For multi-vector probes ($K > 1$), we learn the ordered set $\{\mathbf{v}_\ell^{(0)}, \ldots, \mathbf{v}_\ell^{(K-1)}\}$ sequentially at each controlled layer. The first vector is trained to convergence and then retained as a fixed reference. At each subsequent stage k, *only the current vector* $\mathbf{v}_\ell^{(k)}$ is injected and optimized; previously learned vectors remain inactive in the forward pass and are used solely to define the orthogonality constraint. After each optimizer step, the current vector is reprojected onto the orthogonal complement of the previously learned directions. Training advances to the next vector once the current-stage loss plateaus. Each direction is also evaluated independently, with only the corresponding vector active at each controlled layer. Consequently, each stage optimizes $|\mathcal{S}|d$ parameters, while the complete collection contains $K|\mathcal{S}|d$ scalars; increasing K probes additional independently effective directions rather than simultaneously summing constant offsets.

Vector steering requires a substantially larger learning rate than the other two configurations; the values we use are reported in Table 2.

The auxiliary reference-KL coefficient in Table 2 refers to additional regularization toward π_{ref}, not to the main distillation objective. In particular, a zero auxiliary coefficient does not remove the reverse-KL teacher supervision in OPD or the teacher-trajectory negative log-likelihood in Off-PD.

Prompts. All Qwen3 models are run and trained in *non-thinking* mode (`enable_thinking=False` passed to the chat template) so that the response distribution is directly comparable between the base model, the RL teacher, and every steered student; DeepSeek-R1-Distill students keep thinking mode enabled, as their base model is trained that way. We always use each model's own built-in chat template rather than a hand-written wrapper, and do not manually prepend an additional system message. Each domain uses its own answer-format instruction, which is also what the corresponding verifier parses. The examples below show *user-message content only*, with `{question}` denoting the raw problem statement. Role delimiters and model-specific thinking prefixes are supplied by the native chat template rather than manually inserted into the user message.

For **mathematical reasoning**, the answer-extraction instruction is appended to the question:

Table 2: Learning rates and paradigm-specific hyperparameters.

	RL	**OPD**	**Off-PD**
LR, full fine-tuning	1×10^{-6}	1×10^{-5}	5×10^{-5}
LR, LoRA (r=8)	3×10^{-5}	5×10^{-6}	1×10^{-5}
LR, vector steering	—	1×10^{-1}	5×10^{-2}
Max prompt / response	2,048 / 20,480	3,072 / 16,384	3,072 / 16,384
Prompt batch	128	1,024	1,024
Mini-batch (per step)	32	1,024	1,024
Training samples per prompt n	16	1	1
Evaluation samples per prompt	4	4	4
Evaluation metric	`mean@4`	`mean@4`	`mean@4`
Rollout source	student	student	frozen teacher
Rollout IS correction	—	token, threshold 5.0	off
Auxiliary reference-KL coeff.	0.001 (GRPO) / 0 (DAPO)	0	0
Epochs	1000	1000	1000
Precision	`bfloat16`	`bfloat16`	`bfloat16`

```
{question}
Please reason step by step, and put your final answer within \boxed{}.
```

For **scientific reasoning**, the instruction is placed first and the question is a multiple-choice item whose options are enumerated A–J, so that the verifier can match a single letter:

```
Solve the following problem. Make sure to put the answer (and only answer)
inside \boxed{}.

{question}

A: {option_A}
B: {option_B}
...
J: {option_J}
```

For **code generation**, the instruction requires a single fenced Python block, which the sandbox extracts verbatim and executes against the hidden unit tests:

```
{question}

Write Python code to solve the problem. Present the code in
```python
Your code
```
at the end.
You need to think first then write the Python code.
```

For **instruction following**, the constraints are stated in natural language at the head of the prompt and no separate reasoning instruction is added, since the verifiable constraints concern the response's surface form:

```
Please follow these instructions in your response: {constraint_1} and
{constraint_2} and ... {constraint_n}. {question}
```

Teacher rollouts for off-policy distillation are generated with the *same* model-specific prompt template as the corresponding on-policy runs. Prompt construction is therefore matched, while the two paradigms retain the distinct trajectory-sampling distributions and distillation objectives defined in Section B.1.

Infrastructure. All training runs use 8× or 32× H20 96GB GPUs with FSDP for the actor and vLLM for generation. Vector-steering runs disable CUDA graph capture (`enforce_eager=true`) because the steering hook mutates hidden states inside the vLLM forward pass; LoRA and full-parameter runs leave CUDA graphs enabled. Rollout tensor parallelism is 4 for RL and 2 for distillation. When a fixed model is used for scoring, its parameters are offloaded to CPU between forward passes; this model serves as the reference policy for RL regularization or the teacher for distillation supervision, depending on the paradigm. The two commands below give the on-policy and off-policy distillation invocations for vector steering. The full-parameter and LoRA variants

use the corresponding update parameterizations and hyperparameters in Table 2, with the configuration-specific generation-engine settings described above.

On-Policy Distillation Command (vector steering)

```
# examples/representation/4bopd_single.sh -> 4bopd.sh
python3 -m verl.trainer.main_ppo \
    algorithm.adv_estimator=grpo \
    algorithm.use_kl_in_reward=False \
    algorithm.rollout_correction.rollout_is=token \
    algorithm.rollout_correction.rollout_is_threshold=5.0 \
    algorithm.rollout_correction.rollout_rs=null \
    algorithm.rollout_correction.bypass_mode=false \
    data.train_files=/path/to/DeepMath-103K/train_filtered_level6.parquet \
    data.val_files="['/path/to/AIME2024/test.parquet']" \
    data.train_batch_size=1024 \
    data.max_prompt_length=3072 \
    data.max_response_length=16384 \
    data.filter_overlong_prompts=True \
    data.truncation='error' \
    data.shuffle=True \
    data.seed=42 \
    data.return_raw_chat=True \
    +data.apply_chat_template_kwargs.enable_thinking=False \
    actor_rollout_ref.model.path=$MODEL_PATH \
    +actor_rollout_ref.model.base_model_path=$MODEL_PATH \
    +actor_rollout_ref.ref.model.path=$TEACHER_MODEL_PATH \
    actor_rollout_ref.model.lora_rank=0 \
    actor_rollout_ref.model.enable_trainable_token_vector=true \
    actor_rollout_ref.model.trainable_token_vector_mode=single \
    actor_rollout_ref.model.trainable_token_vector_num=1 \
    actor_rollout_ref.model.trainable_token_vector_layer_start=5 \
    actor_rollout_ref.model.trainable_token_vector_layer_end=20 \
    actor_rollout_ref.model.trainable_token_vector_sampling_method=hypersphere \
    actor_rollout_ref.model.trainable_token_vector_scale=0.1 \
    actor_rollout_ref.model.trainable_token_vector_learnable_alpha=false \
    actor_rollout_ref.model.trainable_token_vector_alpha_init=0.0 \
    actor_rollout_ref.model.trainable_token_vector_curriculum=none \
    actor_rollout_ref.model.trainable_token_vector_force_all_tokens=true \
    actor_rollout_ref.model.use_remove_padding=True \
    actor_rollout_ref.model.enable_gradient_checkpointing=True \
    actor_rollout_ref.actor.optim.lr=1e-1 \
    actor_rollout_ref.actor.optim.lr_warmup_steps_ratio=0.0 \
    actor_rollout_ref.actor.policy_loss.only_reverse_kl_advantages=True \
    actor_rollout_ref.actor.ppo_mini_batch_size=1024 \
    actor_rollout_ref.actor.ppo_micro_batch_size_per_gpu=1 \
    actor_rollout_ref.actor.use_kl_loss=True \
    actor_rollout_ref.actor.kl_loss_coef=0 \
    actor_rollout_ref.actor.kl_loss_type=low_var_kl \
    actor_rollout_ref.actor.entropy_coeff=0 \
    actor_rollout_ref.actor.ppo_max_token_len_per_gpu=32768 \
    actor_rollout_ref.actor.fsdp_config.dtype=bfloat16 \
    actor_rollout_ref.actor.fsdp_config.model_dtype=fp32 \
    actor_rollout_ref.actor.fsdp_config.use_orig_params=True \
    actor_rollout_ref.rollout.name=vllm \
    actor_rollout_ref.rollout.calculate_log_probs=true \
    actor_rollout_ref.rollout.n=1 \
    actor_rollout_ref.rollout.temperature=1.0 \
    actor_rollout_ref.rollout.top_p=1.0 \
    actor_rollout_ref.rollout.tensor_model_parallel_size=2 \
    actor_rollout_ref.rollout.gpu_memory_utilization=0.8 \
    actor_rollout_ref.rollout.enforce_eager=true \
    actor_rollout_ref.rollout.free_cache_engine=True \
    actor_rollout_ref.rollout.max_num_batched_tokens=32768 \
    actor_rollout_ref.rollout.val_kwargs.do_sample=True \
    actor_rollout_ref.rollout.val_kwargs.temperature=1.0 \
    actor_rollout_ref.rollout.val_kwargs.top_p=1.0 \
    actor_rollout_ref.rollout.val_kwargs.n=4 \
    actor_rollout_ref.ref.log_prob_micro_batch_size_per_gpu=1 \
```

```
    actor_rollout_ref.ref.fsdp_config.param_offload=True \
    reward_model.reward_manager=naive \
    trainer.save_vector=true \
    trainer.save_vector_dir=/path/to/trainable_vectors \
    trainer.logger='["console","wandb"]' \
    trainer.n_gpus_per_node=8 \
    trainer.nnodes=1 \
    trainer.test_freq=10 \
    trainer.total_epochs=1000 $@
```

Off-Policy Distillation Command (vector steering)

```
# Stage 1: dump teacher rollouts once
python3 -m verl.trainer.main_generation \
    model.path=$TEACHER_MODEL_PATH \
    data.path=/path/to/DeepMath-103K/train_filtered_level6.parquet \
    data.prompt_key=prompt \
    data.n_samples=1 \
    data.batch_size=1024 \
    rollout.temperature=1.0 \
    rollout.top_p=1.0 \
    rollout.prompt_length=3072 \
    rollout.response_length=16384 \
    rollout.max_num_batched_tokens=32768 \
    rollout.tensor_model_parallel_size=2 \
    data.output_path=$OUTPUT_BASE/round_${i}.parquet

python3 examples/representation/merge_rollouts_to_sft.py \
    --inputs $OUTPUT_BASE/round_*.parquet \
    --output $OUTPUT_BASE/teacher_sft_all.parquet \
    --prompt_key prompt --response_key response

# Stage 2: distill off-policy on the frozen corpus
python3 -m verl.trainer.main_ppo \
    algorithm.adv_estimator=grpo \
    algorithm.rollout_correction.rollout_is=null \
    algorithm.use_kl_in_reward=False \
    data.train_files=$OUTPUT_BASE/teacher_sft_all.parquet \
    data.val_files="['/path/to/AIME2024/test.parquet']" \
    data.train_batch_size=1024 \
    data.max_prompt_length=3072 \
    data.max_response_length=16384 \
    data.return_raw_chat=True \
    +data.apply_chat_template_kwargs.enable_thinking=False \
    actor_rollout_ref.model.path=$MODEL_PATH \
    +actor_rollout_ref.ref.model.path=$TEACHER_MODEL_PATH \
    actor_rollout_ref.model.lora_rank=0 \
    actor_rollout_ref.model.enable_trainable_token_vector=true \
    actor_rollout_ref.model.trainable_token_vector_mode=single \
    actor_rollout_ref.model.trainable_token_vector_num=1 \
    actor_rollout_ref.model.trainable_token_vector_layer_start=5 \
    actor_rollout_ref.model.trainable_token_vector_layer_end=20 \
    actor_rollout_ref.model.trainable_token_vector_sampling_method=hypersphere \
    actor_rollout_ref.model.trainable_token_vector_scale=0.1 \
    actor_rollout_ref.model.trainable_token_vector_learnable_alpha=false \
    actor_rollout_ref.model.trainable_token_vector_alpha_init=0.0 \
    actor_rollout_ref.model.trainable_token_vector_curriculum=none \
    actor_rollout_ref.model.trainable_token_vector_force_all_tokens=true \
    actor_rollout_ref.actor.optim.lr=5e-2 \
    actor_rollout_ref.actor.policy_loss.only_reverse_kl_advantages=true \
    actor_rollout_ref.actor.ppo_mini_batch_size=1024 \
    actor_rollout_ref.actor.ppo_micro_batch_size_per_gpu=1 \
    actor_rollout_ref.actor.ppo_max_token_len_per_gpu=32768 \
    actor_rollout_ref.actor.use_kl_loss=True \
    actor_rollout_ref.actor.kl_loss_coef=0 \
    actor_rollout_ref.actor.entropy_coeff=0 \
    actor_rollout_ref.rollout.offline_teacher_rollout=true \
    actor_rollout_ref.rollout.offline_teacher_data_path=$./teacher_sft_all.parquet \
    actor_rollout_ref.rollout.offline_teacher_response_key=response \
```

```
    actor_rollout_ref.rollout.name=vllm \
    actor_rollout_ref.rollout.n=1 \
    actor_rollout_ref.rollout.tensor_model_parallel_size=2 \
    actor_rollout_ref.rollout.gpu_memory_utilization=0.5 \
    actor_rollout_ref.rollout.enforce_eager=true \
    actor_rollout_ref.ref.fsdp_config.param_offload=True \
    reward_model.reward_manager=naive \
    trainer.save_vector=true \
    trainer.logger='["console","wandb"]' \
    trainer.n_gpus_per_node=8 \
    trainer.nnodes=1 \
    trainer.test_freq=5 \
    trainer.total_epochs=1000 $@
```

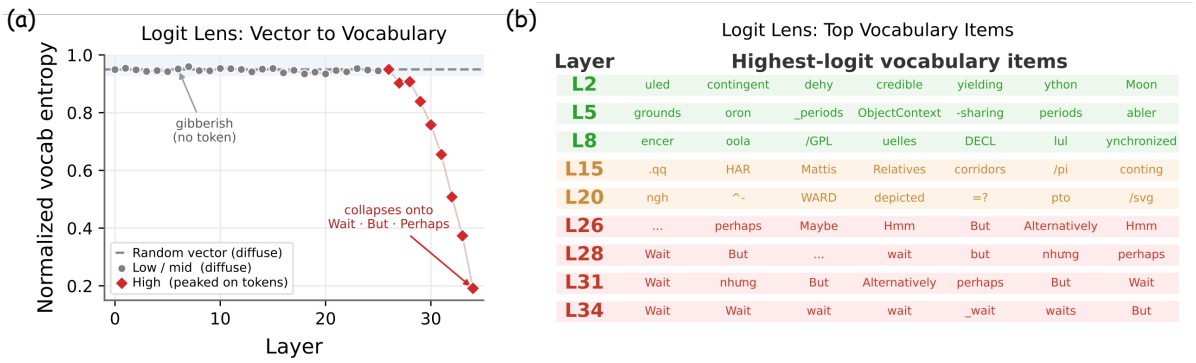

Figure 9: Vocabulary-space readout of steering vectors across network depth. (a) Normalized entropy of $\mathrm{softmax}(\mathrm{RMSNorm}(\mathbf{v}_L)\mathbf{W}_U^\top)$. The dashed line at 0.9498 denotes the mean entropy of normalized Gaussian directions passed through the same readout pipeline. (b) Highest-logit vocabulary items for steering vectors learned at representative layers. All vectors are normalized before readout.

C Additional Experiments for Effective Manifold Capacity

C.1 Vocabulary-Space Readout

The layer-wise experiments in the main text show that fixed-vector steering recovers most of the RL-induced performance gain at lower and intermediate layers, but becomes substantially less effective near the output. To better understand this difference, we examine the vocabulary-space readout of steering vectors learned at different depths. This analysis characterizes how the learned directions align with output vocabulary items and how this alignment changes with injection depth.

For a steering vector $\mathbf{v}_L \in \mathbb{R}^H$ learned at layer L, we bypass subsequent Transformer layers and directly apply RMS normalization followed by the unembedding matrix (Geva et al., 2022):

$$\boldsymbol{\ell}(\mathbf{v}_L) = \mathbf{W}_U \,\mathrm{RMSNorm}(\mathbf{v}_L), \qquad \mathrm{RMSNorm}(\mathbf{v}) = \frac{\mathbf{v}}{\sqrt{\frac{1}{H}\sum_{j=1}^{H} v_j^2 + \varepsilon}}, \qquad \varepsilon = 10^{-6}, \tag{19}$$

where $\mathbf{W}_U \in \mathbb{R}^{V\times H}$ is the unembedding matrix, and H and V denote the hidden dimension and vocabulary size, respectively. The resulting readout distribution is

$$\mathbf{p}_L = \mathrm{softmax}\left(\boldsymbol{\ell}(\mathbf{v}_L)\right). \tag{20}$$

The learned vector norms vary across injection depths. RMS normalization reduces the influence of these magnitude differences, allowing the readout to primarily characterize the vocabulary alignment of each steering direction.

We quantify the concentration of the readout distribution using normalized entropy:

$$\widetilde{\mathcal{H}}(\mathbf{p}_L) = -\frac{1}{\log V}\sum_{t=1}^{V} p_{L,t}\log p_{L,t}. \tag{21}$$

Values closer to one indicate a more diffuse distribution, whereas lower values indicate greater concentration on a smaller set of vocabulary items.

To establish a random-direction baseline, we sample isotropic Gaussian vectors with the same dimensionality as the steering vectors, normalize them to unit L_2 norm, and pass them through the same RMS normalization, unembedding, and softmax operations:

$$\mathbf{z}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_H), \qquad \bar{\mathbf{z}}^{(m)} = \frac{\mathbf{z}^{(m)}}{\|\mathbf{z}^{(m)}\|_2}, \tag{22}$$

$$\mathbf{p}_{\mathrm{null}}^{(m)} = \mathrm{softmax}\left(\mathbf{W}_U \,\mathrm{RMSNorm}(\bar{\mathbf{z}}^{(m)})\right). \tag{23}$$

Averaging the normalized entropy over the M sampled directions gives

$$\widetilde{\mathcal{H}}_{\mathrm{null}} = \frac{1}{M}\sum_{m=1}^{M} \widetilde{\mathcal{H}}\left(\mathbf{p}_{\mathrm{null}}^{(m)}\right) = 0.9498. \tag{24}$$

As shown in Fig. 9 (a), the normalized entropy remains close to the random baseline at lower and intermediate layers, indicating a diffuse vocabulary readout. The highest-logit items from L2 to L20 in Fig. 9(b) likewise consist

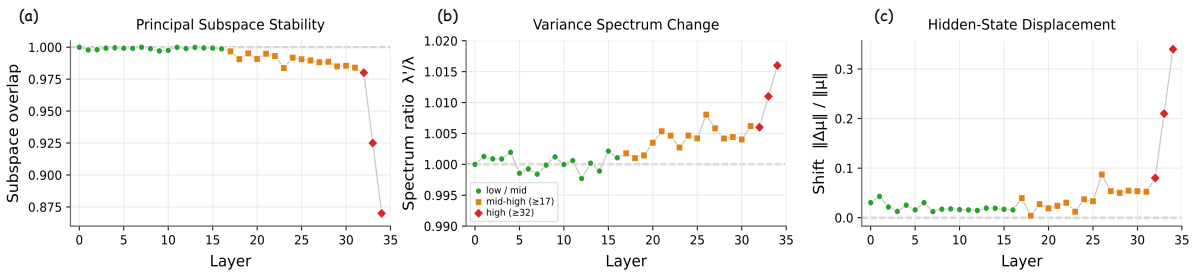

Figure 10: Effect of single-vector steering on hidden-state geometry in Qwen3-4B ($k = 20$, $H = 2560$), measured at layer $L + 1$ for injection at layer L. (a) Mean principal-subspace overlap between clean and steered activations. (b) Ratio of the top-k explained variance after and before steering. (c) Relative displacement of the activation centroid. Dashed lines denote the corresponding reference values.

mainly of heterogeneous subword fragments and symbols, without an apparent common semantic theme. Combined with the strong performance of steering at these depths, this observation shows that effective intermediate-layer interventions need not exhibit a concentrated direct preference for particular output tokens. This pattern is consistent with the role of downstream computation: as a shared additive correction propagates through the remaining layers, its interaction with different hidden states can produce context-dependent output adjustments.

Near the output, the readout becomes increasingly concentrated. Beginning around L26, normalized entropy falls below the random baseline, while the highest-logit items increasingly include `Wait`, `But`, `perhaps`, and `Alternatively`. These tokens are associated with contrast, qualification, and reconsideration, suggesting that high-layer steering directions have more explicit vocabulary-level associations. The readout thus shifts from a diffuse pattern at lower and intermediate layers toward a concentrated set of discourse-related items near the output.

This increase in vocabulary interpretability does not coincide with better steering performance. Instead, the main-text depth sweep shows that fixed-vector steering deteriorates at these higher layers. A direction can therefore have a clear vocabulary-level association while remaining insufficient for effective task adaptation. The issue is not simply whether a direction aligns with meaningful output tokens, but how that direction is applied across token positions. Expressions of contrast, qualification, or reconsideration are useful in different contexts, whereas a fixed steering vector applies the same additive correction at every position, without explicitly adjusting its strength to the current representation.

Summary. The vocabulary readout provides a complementary view of the layer-wise results. Lower- and intermediate-layer steering remains effective despite its diffuse direct readout, whereas high-layer directions exhibit more concentrated vocabulary associations alongside weaker fixed-vector performance. This pattern suggests that direct vocabulary alignment alone does not account for steering effectiveness. One possible explanation is that the remaining layers allow earlier interventions to acquire context-dependent effects, while later interventions have less opportunity for such transformation. The gains from high-layer gating are consistent with this account, supporting input dependence as a useful factor in understanding the depth sensitivity of fixed-vector steering.

C.2 Local Representation Geometry under Single-Vector Steering

Section 2.3 shows that single-vector steering recovers most of the RL-induced performance gain at lower and intermediate layers, but becomes less effective near the output. The training dynamics and gating results suggest that the expressiveness of a fixed additive correction contributes to this depth dependence. Here, we complement the performance analysis by examining changes in hidden-state geometry across injection depths, focusing on the leading principal subspaces, their associated variance, and the activation means.

Experimental Setup. For each injection layer L, we collect hidden states from the base and steered models on the same held-out inputs. The steering vector is added to the output of layer L, and representations are measured at the output of layer $L + 1$. This captures the local effect of steering after one subsequent Transformer layer. Aggregating all valid token positions gives

$$\mathbf{X}_L \in \mathbb{R}^{N\times H}, \qquad \mathbf{X}'_L \in \mathbb{R}^{N\times H}, \tag{25}$$

where N is the number of token positions and H is the hidden dimension. The subscript L indexes the injection layer; both matrices contain activations measured at layer $L + 1$. Their respective means are

$$\boldsymbol{\mu}_L = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{L,i}, \qquad \boldsymbol{\mu}'_L = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}'_{L,i}. \tag{26}$$

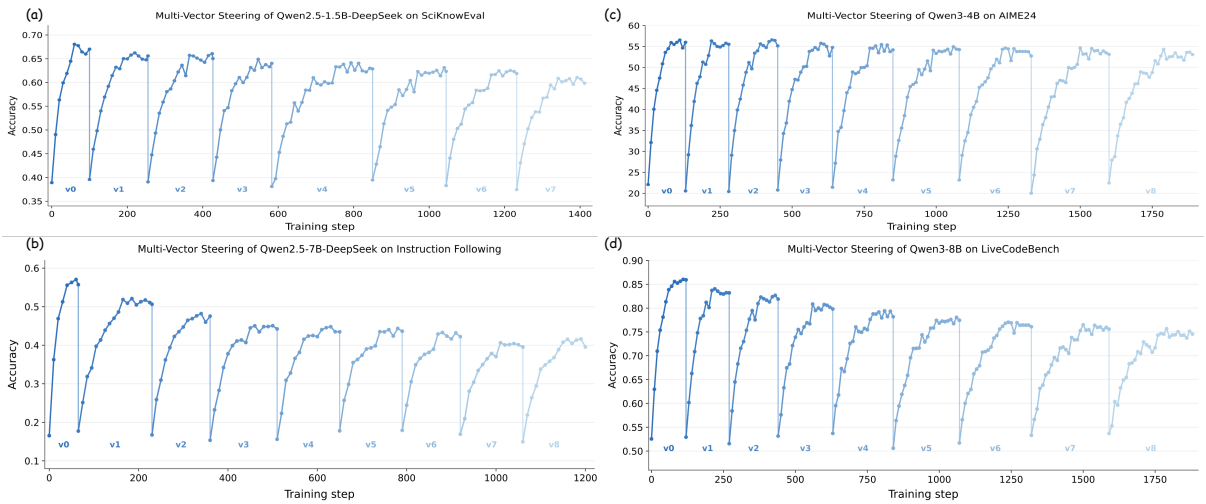

Figure 11: Multi-vector steering across all four task domains and model scales. The first vector is consistently the most effective, with subsequent directions showing gradually decreasing standalone performance; the decay is slower for larger models.

We center each activation matrix using its own mean and independently compute low-rank singular value decompositions:

$$\mathbf{X}_{L,c} \approx \mathbf{U}_L \mathbf{\Sigma}_L \mathbf{V}_L^\top, \qquad \mathbf{X}'_{L,c} \approx \mathbf{U}'_L \mathbf{\Sigma}'_L \mathbf{V}'_L{}^\top. \tag{27}$$

The first k right singular vectors form the orthonormal principal-component bases $\mathbf{V}_{L,k}, \mathbf{V}'_{L,k} \in \mathbb{R}^{H\times k}$. Writing $\Sigma_{L,j}$ and $\Sigma'_{L,j}$ for the corresponding singular values, the covariance eigenvalues are

$$\lambda_{L,j} = \frac{\Sigma_{L,j}^2}{N-1}, \qquad \lambda'_{L,j} = \frac{\Sigma'_{L,j}{}^2}{N-1}. \tag{28}$$

For Qwen3-4B, we use $k = 20$ and $H = 2560$.

Principal-Subspace Alignment. We measure the alignment between the clean and steered top-k principal subspaces as

$$\text{Overlap}_{L,k} = \frac{1}{k} \sum_{j=1}^{k} \sigma_j \left(\mathbf{V}_{L,k}^\top \mathbf{V}'_{L,k} \right), \tag{29}$$

where $\sigma_j(\cdot)$ denotes the j-th singular value. These singular values are the cosines of the principal angles between the two subspaces, so an overlap close to one indicates similar subspace orientations.

As shown in Fig. 10(a), the overlap remains close to one for lower- and intermediate-layer injections. At these depths, the leading directions of activation variation therefore remain largely aligned after one layer of propagation. The overlap decreases toward the output, with the largest departures occurring at the latest injection sites. This indicates greater changes in the orientation of the leading subspace for these high-layer interventions.

Variance Captured by the Leading Components. We next compare the total variance captured by the top-k components of the steered and clean activations:

$$\rho_{L,k} = \frac{\sum_{j=1}^{k} \lambda'_{L,j}}{\sum_{j=1}^{k} \lambda_{L,j}}. \tag{30}$$

The numerator and denominator use the leading eigenvalues of their respective covariance matrices. A ratio close to one indicates similar amounts of variance captured by the two principal subspaces.

Fig. 10(b) shows that this ratio stays close to one at lower and intermediate injection depths. Along with the high subspace overlap, this suggests that steering at these depths introduces relatively small changes in both the orientation and the total variance of the leading components. The ratio departs further from one near the output, although the changes in captured variance remain modest. Thus, the reduced subspace alignment at the latest injection sites is accompanied by a comparatively small change in the variance captured by the leading components.

Displacement of the Activation Centroid. To characterize changes in the mean representation, we measure the relative centroid displacement:

$$D_L = \frac{\|\boldsymbol{\mu}'_L - \boldsymbol{\mu}_L\|_2}{\|\boldsymbol{\mu}_L\|_2 + 10^{-9}}. \tag{31}$$

This statistic expresses the change in the activation mean relative to the norm of the clean mean.

As shown in Fig. 10 (c), relative centroid displacement is small for lower- and intermediate-layer injections and increases near the output, particularly at the latest injection sites. The strong performance of earlier-layer steering therefore coincides with only a small relative change in the mean representation at the next layer. High-layer steering produces a larger relative mean shift, despite its weaker task performance.

Relation to the Layer-Wise Expressiveness Bottleneck. Effective steering at lower and intermediate layers is accompanied by limited changes in the measured local geometry. Near the output, reduced subspace alignment and larger centroid shifts coexist with weaker task performance. These observations are consistent with the expressiveness analysis in Section 2.3: high-layer degradation occurs despite measurable changes in the representations, and the gains from gating suggest that input-dependent modulation helps make these interventions more effective.

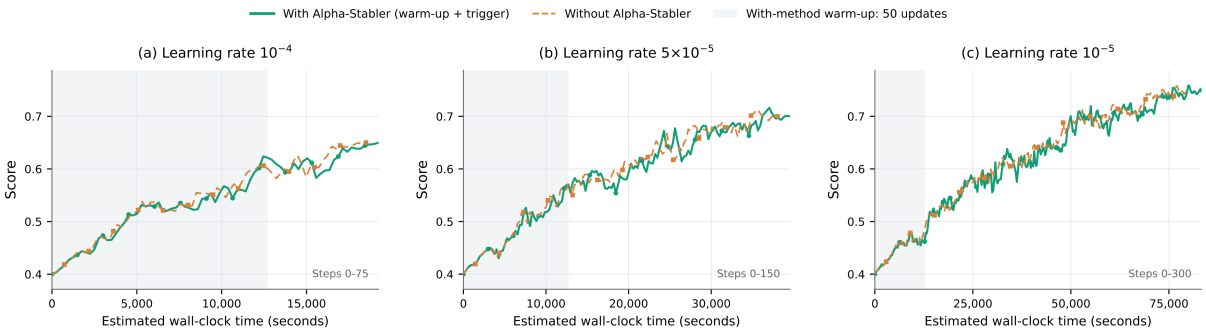

Figure 12: Comparison of pre-collapse score trajectories with and without Alpha-Stabler at learning rates of **(a)** 10^{-4}, **(b)** 5×10^{-5}, and **(c)** 10^{-5}. Shaded regions denote the first 50 warm-up updates of Alpha-Stabler. Controlled and uncontrolled runs exhibit essentially the same score growth before collapse, indicating that Alpha-Stabler's stabilization introduces almost no additional computational overhead.

Algorithm 1 Alpha-Stabler: Warm-Up Calibration and PSI-Triggered Gradient Projection

Require: Actor θ; frozen base θ_0; monitored layers $\mathcal{S}$ at $1/4$, $1/2$, $3/4$ depth
Require: Principal fraction $q = 0.10$; warm-up $T_{\text{warm}} = 50$; interval $M = 3$; EMA $\beta = 0.95$; persistence $p = 3$; thresholds $c_{\text{on}} = 3$, $c_{\text{safe}} = 2$
Require: RL optimizer and LR η_θ; stabilizer $\epsilon > 0$; min. avg. squared token shift $\eta_{\min} > 0$
Ensure: Trained actor θ (inference uses the actor alone)

Run T_{warm} unprojected RL updates; accumulate base activation covariance on actor rollouts
Every M updates, retain detached shifts in $\mathcal{X}_\ell$ if $N > 0$ and $\|X_\ell\|_F^2/N > \eta_{\min}$
Estimate $U_{\ell,r}$ from the reference covariance, $r = \lceil qd \rceil$; freeze bases
for each $\ell \in \mathcal{S}$ **do**
    $\mathcal{W}_\ell \leftarrow \{\|XU_{\ell,r}\|_F^2/(\|X\|_F^2 + \epsilon) : X \in \mathcal{X}_\ell\}$
    if $\mathcal{W}_\ell$ is empty **then**
        Stop and recalibrate
    end if
    $m_\ell \leftarrow \text{median}\, \mathcal{W}_\ell$; $s_\ell \leftarrow 1.4826\ \text{MAD}(\mathcal{W}_\ell) + \epsilon$
    $\tau_\ell^{\text{on}} \leftarrow m_\ell + c_{\text{on}} s_\ell$; $\tau_\ell^{\text{safe}} \leftarrow m_\ell + c_{\text{safe}} s_\ell$
    if $0 < \tau_\ell^{\text{safe}} < \tau_\ell^{\text{on}} < 1$ fails **then**
        Stop and recalibrate
    end if
    $a_\ell \leftarrow 0$; $n_\ell \leftarrow 0$; $\bar{\psi}_\ell \leftarrow m_\ell$
end for
Discard calibration records; $t \leftarrow T_{\text{warm}}$
while training not complete **do**
    Sample rollouts; compute rewards, advantages, and behavior log-probs
    for each minibatch $\mathcal{B}$ (one optimizer update) **do**
        $t \leftarrow t + 1$; clear parameter gradients
        Actor forward for $\mathcal{L}_{\text{RL}}$; retain graph
        if $t \bmod M = 0$ **then**
            Frozen-base activations on the same tokens; $N \leftarrow$ valid positions
            for each $\ell \in \mathcal{S}$ **do**
                $X_\ell \leftarrow \text{sg}(H_\ell(\theta; \mathcal{B}) - H_\ell(\theta_0; \mathcal{B}))$
                if $N > 0$ and $\|X_\ell\|_F^2/N > \eta_{\min}$ **then**
                    $\psi_\ell \leftarrow \|X_\ell U_{\ell,r}\|_F^2/(\|X_\ell\|_F^2 + \epsilon)$
                    $\bar{\psi}_\ell \leftarrow \beta\bar{\psi}_\ell + (1-\beta)\psi_\ell$
                    $n_\ell \leftarrow n_\ell + 1$ if $\bar{\psi}_\ell > \tau_\ell^{\text{on}}$, else 0
                    $a_\ell \leftarrow 1$ if $n_\ell \geq p$; $a_\ell \leftarrow 0$ if $\bar{\psi}_\ell < \tau_\ell^{\text{safe}}$; else keep a_ℓ
                else
                    $n_\ell \leftarrow 0$ ▷ keep $\bar{\psi}_\ell$, a_ℓ
                end if
            end for
        end if
        Fix flags during backward and optimizer step
        Backprop $\mathcal{L}_{\text{RL}}$ with $\widetilde{G}_\ell^{(t)} = G_\ell^{(t)} - a_\ell^{(t)}(G_\ell^{(t)} U_{\ell,r}) U_{\ell,r}^\top$
        Update θ with LR η_θ
    end for
end while
return θ

Note: $H_\ell(\theta; \mathcal{B})$ stacks layer-ℓ activations at valid tokens in $\mathcal{B}$; sg is stop-gradient.

D THEORY OF LOW-DIMENSIONAL RLVR STEERING

We develop a conditional theoretical account of when RLVR-induced output changes admit a compact activation-space intervention. We study recoverability within a specified intervention class, rather than the intrinsic dimension of the full RL parameter or activation trajectory. **Part I** formulates local distillation as Fisher-weighted least squares in Section D.2, gives sufficient conditions for shared teacher-induced shifts in Section D.3, and derives a recovery identity and a three-constant lower bound in Section D.4. **Part II** characterizes sequentially constrained constant-vector fitting in Section D.5 and distinguishes activation variance from output sensitivity in Section D.6. **Part III** analyzes Alpha-Stabler's fixed-reference PSI statistic and backward gradient projection in Section D.7. Section D.8 summarizes the results and their theoretical or empirical scope. Two relations organize the recovery analysis:

$$\rho = \frac{\mathrm{SNR}}{1+\mathrm{SNR}} \qquad \text{and} \qquad \rho \geq \frac{\kappa}{(1+\eta)(1+\varepsilon)}.$$

Here ε bounds sensitivity-weighted principal energy relative to complement energy, η measures residual energy relative to a shared complement component, and κ measures consistency of its amplitude. The quantity ρ is the fraction of locally representable target energy explained by an optimal constant intervention, as defined in Definition D.2; it is not a task-score gain recovery rate.

INFORMAL PREVIEW

Distillation into a frozen student locally amounts to fitting a hidden-state correction to the teacher's output distribution. Logit linearization and a quadratic KL expansion yield a weighted least-squares problem with context-dependent target $\boldsymbol{\Delta}(x)$ (Proposition D.1). An idealized KL-regularized RL optimum provides a separate account of how shared structure may arise: exponential reward tilting (Lemma D.2) determines next-token probability ratios through continuation-success probabilities (Lemma D.3). Shared success profiles, a small relative logit residual, and a consistent representation pullback then give a shared correction with controlled residual (Lemma D.4, Corollary D.1, and Lemma D.5). These are additional structural conditions, not consequences of reward binarity alone. Under a constant metric, the optimal fixed vector is the mean target shift, and its recovery is high when that mean dominates context-dependent fluctuations (Theorem D.1). Table 3 summarizes the quantities entering the sufficient lower bound.

Table 3: Geometric quantities entering the sufficient recovery bound. Their empirical counterparts require matching shift definitions, distributions, and metrics.

Const.	Name	Meaning	Related analysis
ε	principal intrusion	Upper bound on target principal-to-complement energy in the sensitivity metric	PC-energy measurements under compatible metrics and target definitions
η	complement residual	Residual-to-shared energy within the complement	Context dependence and fixed-vector approximation
κ	amplitude consistency	$1/(1+\mathrm{CV}(a)^2)$	Consistency of the shared target amplitude

Theorem D.2 gives $\rho \geq \kappa/((1+\eta)(1+\varepsilon))$. The bound quantifies recovery under the stated localization and shared-component conditions. These conditions are motivated by the empirical findings; their validity and quantitative values for a particular teacher require a separate justification.

PART I. SINGLE-VECTOR STEERING: CONDITIONS FOR HIGH RECOVERY

Section D.1 introduces the local setting. Sections D.2 and D.3 derive the fitting problem and sufficient conditions for shared teacher structure. Section D.4 quantifies the optimal constant-vector fit.

D.1 SETUP AND NOTATION

We consider a frozen pretrained language model with decoder layers $\{f_\ell\}_{\ell=1}^{L}$, hidden width d, vocabulary size V, and unembedding matrix $U \in \mathbb{R}^{V\times d}$, which may be tied or untied. A *context* $x \sim \mathcal{D}$ consists of a token position and its prefix under a specified distillation distribution $\mathcal{D}$. The base model has hidden states and logits

$$\mathbf{h}_\ell(x) \in \mathbb{R}^d, \qquad \mathbf{z}_0(x) = U\mathbf{h}_L(x) \in \mathbb{R}^V,$$

with $p_0(\cdot \mid x) = \mathrm{softmax}(\mathbf{z}_0(x))$. Here $\mathbf{h}_L(x)$ is the representation supplied to the unembedding, including final normalization. For an intervention before normalization, the downstream Jacobian includes that operation. The teacher induces logits $\mathbf{z}_T(x)$ and distribution $p_T(\cdot \mid x)$. Vector steering adds a context-invariant offset $\boldsymbol{\delta}_\ell \in \mathbb{R}^d$

at each controlled layer $\ell \in \mathcal{S} \subseteq \{1, \dots, L\}$: $\mathbf{h}_\ell(x) \mapsto \mathbf{h}_\ell(x) + \boldsymbol{\delta}_\ell$, with all backbone weights frozen. We first analyze a single injection site and extend to multiple sites in Remark D.2.

For a positive-semidefinite matrix A, write $\|v\|_A^2 = v^\top A v$ and $\langle u, v \rangle_A = u^\top A v$. The notation $\|\cdot\|_2$ denotes the Euclidean vector norm or matrix spectral norm; $\lambda_i(\cdot)$ and $\sigma_i(\cdot)$ denote eigenvalues and singular values in descending order. All logarithms are natural and all vectors are columns. Bold symbols denote named vectors, while unbold subscripts such as s_v denote scalar coordinates; matrices such as J_x, F_x, G_x, and B_x use uppercase notation without requiring boldface. The symbol $\mathbf{1}_V$ denotes the all-ones vector in $\mathbb{R}^V$. Generic arguments such as u and v are understood from their stated spaces. We analyze the expected forward-KL objective

$$\mathcal{L}(\boldsymbol{\delta}) = \mathbb{E}_{x \sim \mathcal{D}} \, \mathrm{KL}\big(p_T(\cdot \mid x) \,\big\|\, p_{\boldsymbol{\delta}}(\cdot \mid x)\big). \tag{32}$$

The context distribution is fixed for each fitting problem. All expected quadratic energies used below are finite.

Let $\mathcal{E} \subseteq \mathbb{R}^d$ be a fixed working subspace with orthonormal basis $T \in \mathbb{R}^{d \times h}$. Interventions and target corrections lie in $\mathcal{E}$. Inverses are always taken for the compressed matrices, such as $T^\top G_x T$, assumed positive definite; positive definiteness on $\mathcal{E}$ does not require G_x to be invertible on all of $\mathbb{R}^d$. Setting $T = I_d$ gives the full-space formulas when their inverses exist. For singular metrics, use a common identifiable subspace with a consistent identification of targets across contexts. For results that decompose vectors using P and Q, additionally assume that $\mathcal{E}$ is invariant under these projectors. This ensures that projected targets, means, and corrections remain in the space on which the stated metric bounds apply.

Objects and their dependence on context.

- **Context-dependent values and operators** include $\mathbf{h}_\ell(x)$, $\mathbf{z}_T(x)$, $\mathbf{t}(x)$, $\boldsymbol{\Delta}(x)$, $a(x)$, $\boldsymbol{\xi}(x)$, and J_x, F_x, G_x, B_x.
- **Context-invariant objects** include trainable vectors $\boldsymbol{\delta}_\ell$ and $\mathbf{v}_\ell^{(k)}$, shared directions $\mathbf{s}$ and $\mathbf{w}$, averaged metric G, mean target shift $\boldsymbol{\mu}$, projectors P, Q, and Π_{vocab}, and the constants in the bounds. A trainable vector changes during optimization but is shared across contexts at each step.
- **Distinct shift objects** are the learned intervention $\boldsymbol{\delta}$, the teacher-induced fitting target $\boldsymbol{\Delta}(x) = B_x \mathbf{t}(x)$, and the actual actor-to-base activation differences collected in $X_\ell^{(t)}$ in Part III. They are not identified with one another.
- **Sequence and training indices** are introduced separately: Section D.3 distinguishes prompts q and complete responses Y from contexts $x = (q, y_{<t})$; Part III uses t for a parameter-update index.

Thus, the local fitting problem approximates a context-dependent target family $\{\boldsymbol{\Delta}(x)\}$ by one constant vector in an output-sensitivity metric.

Assumptions. The local reduction uses Assumption D.1. Constant-metric identities use the exact case of Assumption D.3. Assumption D.2 specifies shared teacher structure, while Assumptions D.4–D.5 supply the additional geometry for the three-constant bound. Part III's projection statements require only their own explicitly stated conditions.

Assumption D.1 (Local / trust-region regime)**.** *Fix a teacher, base model, and context distribution. The KL approximation is used only for interventions in a specified neighborhood of the base state. Write*

$$J_x := \left.\frac{\partial \mathbf{z}(x)}{\partial \mathbf{h}_\ell(x)}\right|_{\mathrm{base}} \in \mathbb{R}^{V \times d}, \qquad \mathbf{z}_{\boldsymbol{\delta}}(x) = \mathbf{z}_0(x) + J_x \boldsymbol{\delta} + \mathbf{e}_z(x, \boldsymbol{\delta}).$$

The Jacobian includes all downstream operations. Assume differentiability with $\|\mathbf{e}_z(x, \boldsymbol{\delta})\|_2 = o(\|\boldsymbol{\delta}\|_2)$, and a bounded third derivative of the log-partition function along the segments between the teacher and intervened logits. For $\mathbf{d}_\delta(x) = J_x \boldsymbol{\delta} - \mathbf{t}(x)$, also assume that the combined network-linearization and KL-expansion error is $o(\|\mathbf{d}_\delta(x)\|_2^2)$ in the local regime under consideration. A sufficient condition is $\mathbf{d}_\delta(x) \to 0$ and $\|\mathbf{e}_z(x, \boldsymbol{\delta})\|_2 = o(\|\mathbf{d}_\delta(x)\|_2)$, with uniform or integrable bounds that justify taking expectations of the remainder. Little-o statements concern this joint local regime; differentiability in $\boldsymbol{\delta}$ alone does not imply small relative error when $J_x \boldsymbol{\delta}$ nearly cancels $\mathbf{t}(x)$.

Assumption D.2 (Shared logit component)**.** *The teacher logit shift $\mathbf{t}(x) = \mathbf{z}_T(x) - \mathbf{z}_0(x) \in \mathbb{R}^V$ admits*

$$\mathbf{t}(x) = \underbrace{a(x)\mathbf{s}}_{\text{shared direction}} + \mathbf{r}_\perp(x), \tag{33}$$

where $\mathbf{s}$ is fixed, $\|\mathbf{s}\|_2 = 1$, $\mathbf{1}_V^\top \mathbf{s} = 0$, and $a(x) \geq 0$. The residual obeys $\langle \mathbf{s}, \mathbf{r}_\perp(x) \rangle_{F_x} = 0$ at every context and

$$\mathbb{E}_x \|\mathbf{r}_\perp(x)\|_{F_x}^2 \leq \eta_{\mathrm{logit}} \, \mathbb{E}_x \|a(x)\mathbf{s}\|_{F_x}^2, \qquad \eta_{\mathrm{logit}} \ll 1. \tag{34}$$

Remark D.3 justifies the centered gauge. Section D.3 provides sufficient success-profile conditions for this decomposition; it is not assumed to hold merely because rewards are binary.

Table 4: Recurring symbols for the distillation analysis. Context dependence is indicated by (x) or a subscript x; Part III separately introduces batched training quantities.

Symbol	Type	Meaning
$x \sim \mathcal{D}$	context	Token position and prefix, $x = (q, y_{<t})$
q, Y	sequence level	Prompt and complete response, including termination
$\mathbf{h}_\ell(x) \in \mathbb{R}^d$	ctx.-dep.	Layer-ℓ hidden state
$\mathbf{z}_0(x), \mathbf{z}_T(x) \in \mathbb{R}^V$	ctx.-dep.	Base and teacher logits
$\mathbf{t}(x) = \mathbf{z}_T(x) - \mathbf{z}_0(x)$	ctx.-dep.	Teacher logit shift modulo $\mathbf{1}_V$; Remark D.3
$F_x \in \mathbb{R}^{V\times V}$	ctx.-dep.	Fisher information at teacher logits; Equation (38)
$J_x \in \mathbb{R}^{V\times d}$	ctx.-dep.	Downstream logit Jacobian at the base injection state
$G_x = J_x^\top F_x J_x$	ctx.-dep.	Pullback sensitivity metric
$B_x = T(T^\top G_x T)^{-1} T^\top J_x^\top F_x$	ctx.-dep.	Least-squares pullback into $\mathcal{E}$
$\mathbf{\Delta}(x) = B_x \mathbf{t}(x)$	ctx.-dep.	Locally best target representation correction
$a(x) \in \mathbb{R}_{\geq 0}$	ctx.-dep.	Amplitude in the specified shared-component decomposition
$\boldsymbol{\xi}(x) \in \mathbb{R}^d$	ctx.-dep.	Representation residual within the complement
$\nu_0(x) \in [0, 1]$	ctx.-dep.	Conditional success probability; Equation (42)
$Z_\beta(x)$	ctx.-dep.	Conditional reward-tilt normalizer
$b(x), a_0(x), \mathbf{e}(x)$	ctx.-dep.	Baseline, modulation, and residual of a success profile
$\boldsymbol{\delta}_\ell, \mathbf{v}_\ell^{(k)} \in \mathbb{R}^d$	invariant	Learned constant steering vectors
$\mathbf{s} \in \mathbb{R}^V, \mathbf{w} \in \mathbb{R}^d$	invariant	Shared logit and representation directions
$G = \mathbb{E}_x[G_x]$	invariant	Averaged sensitivity metric
$\boldsymbol{\mu} = \mathbb{E}_x[\mathbf{\Delta}(x)]$	invariant	Mean target; optimal constant vector when $G_x = G$
$C, P, Q = I - P$	invariant	Reference activation covariance and its spectral projectors
$\Pi_{\text{vocab}} = I - \mathbf{1}_V \mathbf{1}_V^\top / V$	invariant	Vocabulary-centering projector, distinct from P
$c_\beta = e^{1/\beta} - 1$	invariant	Tilt strength for binary rewards
$\varepsilon, \eta, \kappa, \rho$	invariant	Intrusion bound, residual ratio, amplitude consistency, and energy recovery
$\eta_{\text{logit}}, \theta, \zeta_B, \zeta_R$	invariant	Relative logit-residual and pullback bounds

Assumption D.3 (Metric homogeneity)**.** *Let $\bar{G} = \mathbb{E}_{x\sim\mathcal{D}}[G_x]$ and abbreviate $G = \bar{G}$, with $T^\top G T$ positive definite. Exact homogeneous-model identities assume $G_x = G$ as quadratic forms on $\mathcal{E}$. When approximation is invoked, its quantitative form is $(1-\zeta)G \preceq G_x \preceq (1+\zeta)G$ on $\mathcal{E}$, uniformly over the considered contexts, for $0 \leq \zeta < 1$. Remark D.1 treats this perturbation for a fixed target family. Consistency of $B_x\mathbf{s}$ is a separate condition, quantified by ζ_B in Lemma D.5; it does not follow from metric homogeneity.*

Assumption D.4 (Off-principal localization)**.** *Let P be the Euclidean orthogonal projector onto the top-m principal subspace of the reference activation covariance C in Definition D.3, and let $Q = I - P$. Here m is a general retained rank; the monitoring construction in Section D.7 uses the particular rank $r = \lceil 0.10d \rceil$. Assume compatibility with the sensitivity metric, $P^\top G Q = 0$ on the working subspace. Then*

$$\mathbb{E}_x\|\mathbf{\Delta}(x)\|_G^2 = \mathbb{E}_x\|P\mathbf{\Delta}(x)\|_G^2 + \mathbb{E}_x\|Q\mathbf{\Delta}(x)\|_G^2. \tag{35}$$

Assume in addition the target-energy bound

$$\mathbb{E}_x\|P\mathbf{\Delta}(x)\|_G^2 \leq \varepsilon\, \mathbb{E}_x\|Q\mathbf{\Delta}(x)\|_G^2, \qquad 0 \leq \varepsilon \ll 1.$$

The operator condition gives a G-orthogonal split for every vector in $\mathcal{E}$; equality of averaged energies for one distribution is weaker. Neither covariance principal structure nor a low Euclidean PC fraction alone guarantees these sensitivity-weighted conditions. Remark D.1 treats approximate compatibility.

Assumption D.5 (Shared direction in the complement)**.** *There is a fixed $\mathbf{w} \in \mathcal{E} \cap \text{range}(Q)$ with $\|\mathbf{w}\|_G = 1$ such that*

$$Q\mathbf{\Delta}(x) = \underbrace{a(x)\mathbf{w}}_{\text{shared complement component}} + \boldsymbol{\xi}(x), \qquad a(x) \geq 0, \tag{36}$$

where $\boldsymbol{\xi}(x) \in \mathcal{E} \cap \text{range}(Q)$ and

$$\mathbb{E}_x[\boldsymbol{\xi}(x)] = \mathbf{0}, \qquad \mathbb{E}_x[a(x)\langle \mathbf{w}, \boldsymbol{\xi}(x)\rangle_G] = 0. \tag{37}$$

Assume $0 < \mathbb{E}_x[a(x)^2] < \infty$ and finite target energy, and define

$$\kappa := \frac{(\mathbb{E}_x[a(x)])^2}{\mathbb{E}_x[a(x)^2]} \in (0, 1], \qquad \eta := \frac{\mathbb{E}_x\|\boldsymbol{\xi}(x)\|_G^2}{\mathbb{E}_x[a(x)^2]} \geq 0.$$

Then $\kappa = 1$ exactly when $a(x)$ is constant almost surely, and $\kappa = 1/(1 + \mathrm{CV}(a)^2)$ with $\mathrm{CV}(a)^2 = \mathrm{Var}_x(a(x))/(\mathbb{E}_x[a(x)])^2$. Nonnegativity fixes the shared amplitude's sign; η is a residual-energy ratio, not a rank. The two moment conditions in Equation (37) have distinct roles: zero mean identifies the shared mean, while weighted orthogonality removes the mixed energy term. Pointwise orthogonality is sufficient for the latter but is not required. The amplitude here belongs to the representation-space decomposition. It is identified with a fixed positive rescaling of the logit-space amplitude only under the compatible pullback conditions discussed in Remark D.5; otherwise the two decompositions use separately defined amplitudes.

Remark D.1 (Perturbations of the homogeneous model)**.** (i) Varying metrics. *Let $\bar{G}_{\mathcal{E}} = T^\top \bar{G} T$ and $\mathbf{b}_{\mathcal{E}} = T^\top \mathbb{E}_x[G_x \boldsymbol{\Delta}(x)]$. The quadratic optimum is $\boldsymbol{\delta}^\star = T\bar{G}_{\mathcal{E}}^{-1}\mathbf{b}_{\mathcal{E}}$, with explained energy $S = \mathbf{b}_{\mathcal{E}}^\top \bar{G}_{\mathcal{E}}^{-1}\mathbf{b}_{\mathcal{E}}$. For $E_{\mathrm{var}} = \mathbb{E}_x\|\boldsymbol{\Delta}(x)\|^2_{G_x} > 0$, its recovery is $\rho_{\mathrm{var}} = S/E_{\mathrm{var}}$. For a fixed target family and the metric bounds in Assumption D.3, comparison with constant-metric recovery ρ_G yields*

$$\frac{1-\zeta}{1+\zeta}(1-\rho_G) \le 1-\rho_{\mathrm{var}} \le \frac{1+\zeta}{1-\zeta}(1-\rho_G).$$

Indeed, the metric inequalities bound the residual objective for every admissible constant vector and therefore its infimum, as well as the total target energy. Comparing different target families requires an additional target-perturbation argument.

(ii) Approximately orthogonal splits. *Replace exact compatibility by $|\langle u, v\rangle_G| \le \gamma\|u\|_G\|v\|_G$ for $u \in \mathcal{E} \cap \mathrm{range}(P)$ and $v \in \mathcal{E} \cap \mathrm{range}(Q)$, with $0 \le \gamma < 1$. Then*

$$(1-\gamma)(\|Pz\|_G^2 + \|Qz\|_G^2) \le \|z\|_G^2 \le (1+\gamma)(\|Pz\|_G^2 + \|Qz\|_G^2).$$

This follows by expanding the cross term and using $2ab \le a^2 + b^2$. Under the remaining conditions of Theorem D.2, bounding both the mean energy and total target energy gives $\rho \ge \frac{1-\gamma}{1+\gamma}\frac{\kappa}{(1+\eta)(1+\varepsilon)}$.

(iii) Distinct residual and metric conditions. *Corollary D.1 bounds the logit residual, and Lemma D.5 bounds a representation residual relative to a specified direction. Applying these estimates to η additionally requires localization, zero residual mean, and weighted orthogonality as in Assumption D.5, or explicit bounds on their deviations. Corollary D.4 compares Euclidean and sensitivity-weighted intrusion only for the same collection of shifts.*

D.2 Distillation reduces to weighted least squares

The categorical distribution has natural parameter $\mathbf{z} \in \mathbb{R}^V$, log-partition function $A(\mathbf{z}) = \log \sum_{v=1}^V e^{z_v}$, and mean map $\nabla A(\mathbf{z}) = \mathrm{softmax}(\mathbf{z})$. Writing $p = \mathrm{softmax}(\mathbf{z})$, its Hessian is

$$F(\mathbf{z}) = \nabla^2 A(\mathbf{z}) = \mathrm{diag}(p) - pp^\top \succeq 0. \tag{38}$$

Throughout, $F_x = F(\mathbf{z}_T(x))$ is evaluated at teacher logits. It satisfies $F_x \mathbf{1}_V = \mathbf{0}$.

Lemma D.1 (KL as a Bregman divergence and its local quadratic form)**.** *For $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^V$, $\mathrm{KL}(\mathrm{softmax}(\mathbf{z}') \,\|\, \mathrm{softmax}(\mathbf{z})) = D_A(\mathbf{z} \,\|\, \mathbf{z}')$, where D_A is the Bregman divergence generated by A. Under Assumption D.1,*

$$\mathrm{KL}\big(p_T(\cdot \mid x) \,\big\|\, p_{\boldsymbol{\delta}}(\cdot \mid x)\big) = \tfrac{1}{2}\big(J_x\boldsymbol{\delta} - \mathbf{t}(x)\big)^\top F_x \big(J_x\boldsymbol{\delta} - \mathbf{t}(x)\big) + o\big(\|J_x\boldsymbol{\delta} - \mathbf{t}(x)\|_2^2\big). \tag{39}$$

Proof. The identity follows by substituting A and ∇A into D_A. Set $d = J_x\boldsymbol{\delta} - \mathbf{t}(x)$ and $e = \mathbf{e}_z(x, \boldsymbol{\delta})$, so the actual logit difference from the teacher is $d + e$. Taylor expansion at $\mathbf{z}_T(x)$ gives

$$\mathrm{KL}(p_T \| p_{\boldsymbol{\delta}}) = \tfrac{1}{2}d^\top F_x d + d^\top F_x e + \tfrac{1}{2}e^\top F_x e + R_A(d+e).$$

If L_A bounds the third derivative of A in trilinear operator norm along the logit segment, then

$$\big|\mathrm{KL}(p_T \| p_{\boldsymbol{\delta}}) - \tfrac{1}{2}d^\top F_x d\big| \le \|F_x\|_2\|d\|_2\|e\|_2 + \tfrac{1}{2}\|F_x\|_2\|e\|_2^2 + \tfrac{L_A}{6}\|d+e\|_2^3.$$

The combined remainder condition yields Equation (39); the integrability condition permits the corresponding expected expansion. □

Definition D.1 (Representational metric and target shift)**.** *For each context, define*

$$G_x := J_x^\top F_x J_x \in \mathbb{R}^{d\times d}, \qquad B_x := T(T^\top G_x T)^{-1} T^\top J_x^\top F_x \in \mathbb{R}^{d\times V},$$

$$\boldsymbol{\Delta}(x) := \arg\min_{\mathbf{u}\in\mathcal{E}} \|J_x\mathbf{u} - \mathbf{t}(x)\|_{F_x}^2 = B_x\mathbf{t}(x),$$

where $T^\top G_x T$ is positive definite. When $T = I_d$, $B_x = G_x^{-1}J_x^\top F_x$. The metric measures local output sensitivity; $\boldsymbol{\Delta}(x)$ is the best linearized hidden-state correction for reproducing the teacher shift. This target is context-dependent even though the learned intervention is constant. It is a fitting target, not an observed teacher-minus-base hidden-state difference.

Proposition D.1 (Distillation as local weighted least squares)**.** *Under Assumption D.1, the local quadratic surrogate is*

$$\mathcal{L}_{\mathrm{quad}}(\boldsymbol{\delta}) = \tfrac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}(\boldsymbol{\delta} - \boldsymbol{\Delta}(x))^\top G_x(\boldsymbol{\delta} - \boldsymbol{\Delta}(x)) + C_{\mathrm{irr}}, \tag{40}$$

where $C_{\mathrm{irr}} \geq 0$ is independent of $\boldsymbol{\delta}$ and is the residual teacher-logit energy not fitted by the local Jacobian on $\mathcal{E}$, measured in the Fisher seminorm. The KL objective differs by the expected remainder in Lemma D.1.

Proof. Write $\mathbf{t}(x) = J_x\boldsymbol{\Delta}(x) + \mathbf{r}_x$. The least-squares normal equations give $T^\top J_x^\top F_x \mathbf{r}_x = 0$. Since $\boldsymbol{\delta} - \boldsymbol{\Delta}(x) \in \mathcal{E}$,

$$\|J_x\boldsymbol{\delta} - \mathbf{t}(x)\|_{F_x}^2 = \|J_x(\boldsymbol{\delta} - \boldsymbol{\Delta}(x))\|_{F_x}^2 + \|\mathbf{r}_x\|_{F_x}^2.$$

The first term is $\|\boldsymbol{\delta} - \boldsymbol{\Delta}(x)\|_{G_x}^2$. Taking expectations proves Equation (40), with $C_{\mathrm{irr}} = \frac{1}{2}\mathbb{E}_x\|\mathbf{r}_x\|_{F_x}^2$. □

Remark D.2 (Multiple injection sites)**.** *Stack the steering vectors and use the joint Jacobian from all injection sites to the logits. Equation (40) holds with the full $G_x = J_x^\top F_x J_x$, including cross-layer blocks; it is not in general a sum of independent layer-wise fitting problems. Injection at the unembedding input gives $J_x = U$, whereas earlier injection includes normalization and all other downstream operations.*

Local surrogate versus training dynamics. The reduction describes a fixed-distribution local objective. At a fixed context, forward and reverse KL have the same second-order Fisher term around the teacher, but generally differ beyond this local expansion. Changing on-policy prefix distributions or using an approximate token-level training gradient changes the optimization problem. Consequently, Equation (40) is not an identity for the complete OPD training trajectory or for every off-policy surrogate. Moreover, the unconstrained optimum of the quadratic model need not remain in the neighborhood where that model is accurate. A small teacher–student output gap alone does not ensure this: the Jacobian, its conditioning, and the intervention size also matter. Section D.4 makes the distinction between unconstrained surrogate recovery and locally admissible recovery explicit.

D.3 The RLVR teacher shift: exponential tilting and shared structure

We derive the complete-response optimum of an idealized KL-regularized RL objective and then its next-token shifts. The analysis separates sequence rewards, complete-response probability ratios, and token-level corrections. Shared structure requires conditions on continuation-success profiles and on the student's pullback.

Sequence notation and optimization setting. Let q be a prompt, Y a complete response including termination, and $x = (q, y_{<t})$ a next-token context. Responses terminate almost surely under the base policy. Rewards are bounded, $\beta > 0$, and policies are supported on the base policy's support. The symbol $\pi^\star$ denotes the optimizer of this stated objective. Empirical teacher checkpoints instead enter the local fitting analysis through their actual logit shifts $\mathbf{t}(x)$.

Lemma D.2 (KL-regularized RL optimum as an exponential tilt)**.** *For a fixed prompt q, the objective*

$$\max_{\pi}\ \mathbb{E}_{Y\sim\pi(\cdot|q)}[r(q,Y)] - \beta\ \mathrm{KL}\left(\pi(\cdot\mid q)\,\middle\|\,p_0(\cdot\mid q)\right), \qquad \beta > 0,$$

is strictly concave on its finite-KL domain and has the unique maximizer

$$\pi^\star(Y\mid q) = \frac{p_0(Y\mid q)e^{r(q,Y)/\beta}}{Z_\beta(q)}, \qquad Z_\beta(q) = \mathbb{E}_{Y\sim p_0(\cdot|q)}[e^{r(q,Y)/\beta}]. \tag{41}$$

Equivalently, $\log(\pi^\star(Y\mid q)/p_0(Y\mid q)) = r(q,Y)/\beta - \log Z_\beta(q)$ on the support.

Proof. Bounded rewards make $Z_\beta(q)$ positive and finite, and the proposed optimizer has finite KL. Substitution rewrites the objective as $\beta\log Z_\beta(q) - \beta\,\mathrm{KL}(\pi(\cdot\mid q)\|\pi^\star(\cdot\mid q))$. Nonnegativity of KL gives the unique maximum. Strict concavity follows from strict convexity of KL and linearity of the reward term. □

Lemma D.3 (From complete responses to next-token probabilities)**.** *For $x = (q, y_{<t})$ with positive base prefix probability, define*

$$Z_\beta(x) := \mathbb{E}_{Y\sim p_0(\cdot|x)}[e^{r(q,Y)/\beta}], \qquad \nu_0(x) := \Pr_{Y\sim p_0(\cdot|x)}(r(q,Y) = 1), \tag{42}$$

where complete responses extend the prefix. For a supported next token v, write xv for the extended context. Then

$$\pi^\star(v\mid x) = p_0(v\mid x)\frac{Z_\beta(xv)}{Z_\beta(x)}. \tag{43}$$

For binary rewards $r \in \{0,1\}$ and $c_\beta = e^{1/\beta} - 1 > 0$,

$$\log\frac{\pi^\star(v\mid x)}{p_0(v\mid x)} = \log[1 + c_\beta\nu_0(xv)] - \log[1 + c_\beta\nu_0(x)]. \tag{44}$$

At termination the conditional quantities are evaluated on the resulting complete response.

Proof. Summing Equation (41) over responses extending a prefix gives

$$\pi^\star(y_{<t} \mid q) = p_0(y_{<t} \mid q)\frac{Z_\beta(x)}{Z_\beta(q)}.$$

Taking the ratio for $y_{<t}v$ and $y_{<t}$ proves Equation (43). For binary rewards, $Z_\beta(x) = 1 - \nu_0(x) + e^{1/\beta}\nu_0(x) = 1 + c_\beta\nu_0(x)$, which gives Equation (44). □

Thus, relative next-token changes depend on candidate-token continuation-success probabilities, not just the terminal reward of an already observed response. The tower property gives $\nu_0(x) = \sum_v p_0(v \mid x)\nu_0(xv)$. Neither this identity nor binary rewards imposes a common direction across different contexts.

Remark D.3 (Softmax gauge and centered logits)**.** *Since* $\mathrm{softmax}(\mathbf{z} + c\mathbf{1}_V) = \mathrm{softmax}(\mathbf{z})$*, logits are defined modulo the all-ones direction. Choose the centered gauge* $\Pi_{\mathrm{vocab}} = I - \mathbf{1}_V\mathbf{1}_V^\top/V$*. Because* $F_x\mathbf{1}_V = \mathbf{0}$*,*

$$F_x\Pi_{\mathrm{vocab}} = F_x, \qquad \|\Pi_{\mathrm{vocab}}\mathbf{u}\|_{F_x} = \|\mathbf{u}\|_{F_x}, \qquad B_x\Pi_{\mathrm{vocab}}\mathbf{u} = B_x\mathbf{u} \quad \textit{for all } \mathbf{u} \in \mathbb{R}^V. \tag{45}$$

Centering preserves Fisher seminorms, Fisher inner products, and $\mathbf{\Delta}(x)$*. For the idealized teacher with finite logits, we may use* $\mathbf{t}(x) = \Pi_{\mathrm{vocab}}\log(\pi^\star(\cdot \mid x)/p_0(\cdot \mid x))$ *and choose* $\mathbf{1}_V^\top\mathbf{s} = 0$*. The vocabulary-centering projector* Π_{vocab} *is distinct from the activation-principal projector* P*.*

Lemma D.4 (A sufficient condition for a shared logit component)**.** *Let* $\mathbf{s} \in \mathbb{R}^V$ *be fixed,* $\|\mathbf{s}\|_2 = 1$*, and* $\mathbf{1}_V^\top\mathbf{s} = 0$*. Suppose the continuation-success profiles obey*

$$\nu_0(xv) = b(x) + a_0(x)s_v + e_v(x), \qquad b(x) \in [0,1], \quad a_0(x) \geq 0, \quad \nu_0(xv) \in [0,1], \tag{46}$$

where $b(x)$ *is token-independent,* s_v *is the* v*-th scalar coordinate of* $\mathbf{s}$*, and* $\mathbf{e}(x)$ *is a residual profile with coordinates* $e_v(x)$*. Define* $\mathbf{d}(x) = a_0(x)\mathbf{s} + \mathbf{e}(x)$*. Then*

$$\Pi_{\mathrm{vocab}}\log\frac{\pi^\star(\cdot \mid x)}{p_0(\cdot \mid x)} = \widehat{a}(x)\mathbf{s} + \widehat{\mathbf{r}}(x), \qquad \widehat{a}(x) = \frac{c_\beta a_0(x)}{1 + c_\beta b(x)} \geq 0, \tag{47}$$

with coordinatewise logarithms and ratios, and

$$\|\widehat{\mathbf{r}}(x)\|_2 \leq c_\beta\|\mathbf{e}(x)\|_2 + \frac{c_\beta^2}{2}\|\mathbf{d}(x)\|_4^2, \qquad \|\mathbf{d}(x)\|_4^2 = \left(\sum_v |d_v(x)|^4\right)^{1/2}. \tag{48}$$

Proof. For $f(z) = \log(1 + c_\beta z)$ on $[0,1]$, $|f'(z)| \leq c_\beta$ and $|f''(z)| \leq c_\beta^2$. Taylor expansion between $b(x)$ and $b(x) + d_v(x)$, both in that interval, yields

$$f(b(x) + d_v(x)) = f(b(x)) + f'(b(x))d_v(x) + R_v(x), \qquad |R_v(x)| \leq \tfrac{1}{2}c_\beta^2|d_v(x)|^2.$$

Centering Equation (44) removes the token-independent terms. Since $\Pi_{\mathrm{vocab}}\mathbf{s} = \mathbf{s}$, $\widehat{\mathbf{r}}(x) = \Pi_{\mathrm{vocab}}(f'(b(x))\mathbf{e}(x) + R(x))$. The projector is a Euclidean contraction, and $\|R(x)\|_2 \leq \frac{1}{2}c_\beta^2\|\mathbf{d}(x)\|_4^2$. These inequalities prove Equation (48). □

The bound is absolute. Obtaining high relative alignment additionally requires a residual small compared with $\widehat{a}(x)\|\mathbf{s}\|_{F_x}$, as specified next; a small absolute residual alone is insufficient when the shared amplitude is small.

Corollary D.1 (Fisher-orthogonal shared-component decomposition)**.** *Assume Lemma D.4,* $\|\mathbf{s}\|_{F_x} > 0$*, and* $0 \leq \theta < 1$ *such that*

$$\|\widehat{\mathbf{r}}(x)\|_{F_x} \leq \theta\widehat{a}(x)\|\mathbf{s}\|_{F_x} \quad \textit{at every considered context.} \tag{49}$$

For $\mathbf{t}(x) = \widehat{a}(x)\mathbf{s} + \widehat{\mathbf{r}}(x)$*, define*

$$a(x) := \frac{\langle \mathbf{s}, \mathbf{t}(x)\rangle_{F_x}}{\|\mathbf{s}\|_{F_x}^2}, \qquad \mathbf{r}_\perp(x) := \mathbf{t}(x) - a(x)\mathbf{s}. \tag{50}$$

Then $a(x) \geq (1-\theta)\widehat{a}(x) \geq 0$*,* $\langle \mathbf{s}, \mathbf{r}_\perp(x)\rangle_{F_x} = 0$*, and*

$$\mathbb{E}_x\|\mathbf{r}_\perp(x)\|_{F_x}^2 \leq \frac{\theta^2}{(1-\theta)^2}\,\mathbb{E}_x\|a(x)\mathbf{s}\|_{F_x}^2. \tag{51}$$

This meets Assumption D.2's small-residual requirement when θ *is sufficiently small.*

Proof. By Equation (50), $a(x) = \widehat{a}(x) + \langle \mathbf{s}, \widehat{\mathbf{r}}(x)\rangle_{F_x}/\|\mathbf{s}\|_{F_x}^2$. Cauchy–Schwarz bounds the second term in absolute value by $\theta\widehat{a}(x)$. The vector $a(x)\mathbf{s}$ minimizes the Fisher squared distance to $\mathbf{t}(x)$ over the one-dimensional span, so

$$\|\mathbf{r}_\perp(x)\|_{F_x} \leq \|\mathbf{t}(x) - \widehat{a}(x)\mathbf{s}\|_{F_x} \leq \theta\widehat{a}(x)\|\mathbf{s}\|_{F_x} \leq \frac{\theta}{1-\theta}a(x)\|\mathbf{s}\|_{F_x}.$$

Squaring and averaging proves the bound. Orthogonality follows directly from the definition, and Remark D.3 ensures gauge invariance. □

Remark D.4 (Small positive β and uniform residual bounds)**.** *A multiplicative success profile provides a bound without diverging constants as $\beta \downarrow 0$, under an additional uniform success condition. Suppose $\nu_0(xv) \geq \nu_{\min} > 0$, $c_\beta \nu_{\min} \geq 1/\varrho$ for $0 < \varrho \leq 1$, and*

$$\log \nu_0(xv) = \tilde{b}(x) + a_0(x) s_v + \tilde{e}_v(x), \qquad a_0(x) \geq 0. \tag{52}$$

Using $\log(1 + c_\beta z) = \log c_\beta + \log z + \log(1 + 1/(c_\beta z))$ gives

$$\Pi_{\text{vocab}} \log \frac{\pi^\star(\cdot \mid x)}{p_0(\cdot \mid x)} = a_0(x)\mathbf{s} + \Pi_{\text{vocab}} \tilde{\mathbf{e}}(x) + \boldsymbol{\epsilon}(x), \qquad \|\boldsymbol{\epsilon}(x)\|_2 \leq \sqrt{V}\, \varrho.$$

Indeed, each last logarithm lies in $[0, \varrho]$, and centering is a contraction. Corollary D.1 applies only if its relative Fisher-residual condition also holds. The positive lower bound on success probabilities excludes zero-success continuations and is not automatic for binary rewards. This regime concerns a small positive penalty coefficient in an idealized optimum, not a small realized KL divergence or a zero-penalty checkpoint.

Lemma D.5 (Consistency of the representation pullback)**.** *Retain Definition D.1 and Assumption D.2. For fixed nonzero $\mathbf{w}_0 \in \mathcal{E}$, define $\mathbf{e}_h(x) = \boldsymbol{\Delta}(x) - a(x)\mathbf{w}_0$. Then*

$$\mathbf{e}_h(x) = a(x)(B_x \mathbf{s} - \mathbf{w}_0) + B_x \mathbf{r}_\perp(x). \tag{53}$$

Let G be a fixed positive definite metric on $\mathcal{E}$, set $E_0 = \mathbb{E}_x[a(x)^2]\|\mathbf{w}_0\|_G^2 > 0$, and suppose

$$\mathbb{E}_x\left[a(x)^2 \|B_x \mathbf{s} - \mathbf{w}_0\|_G^2\right] \leq \zeta_B E_0, \qquad \mathbb{E}_x \|B_x \mathbf{r}_\perp(x)\|_G^2 \leq \zeta_R E_0 \tag{54}$$

for $\zeta_B, \zeta_R \geq 0$. Then

$$\mathbb{E}_x \|\mathbf{e}_h(x)\|_G^2 \leq 2(\zeta_B + \zeta_R) E_0. \tag{55}$$

The constants separately control variation of the shared pullback and the pulled-back residual energy.

Proof. Substitute $\mathbf{t}(x) = a(x)\mathbf{s} + \mathbf{r}_\perp(x)$ into $\boldsymbol{\Delta}(x) = B_x \mathbf{t}(x)$ and subtract $a(x)\mathbf{w}_0$ to obtain Equation (53). Applying $\|u + v\|_G^2 \leq 2\|u\|_G^2 + 2\|v\|_G^2$ and taking expectations proves Equation (55). □

Under exact homogeneity, define $A_x = F_x^{1/2} J_x T$ and $\mathcal{P}_x = A_x (A_x^\top A_x)^{-1} A_x^\top$. This is an orthogonal projector, so

$$\|B_x \mathbf{r}_\perp(x)\|_G^2 = (F_x^{1/2} \mathbf{r}_\perp(x))^\top \mathcal{P}_x (F_x^{1/2} \mathbf{r}_\perp(x)) \leq \|\mathbf{r}_\perp(x)\|_{F_x}^2. \tag{56}$$

Consequently, the smallest admissible residual ratio in Equation (54) is at most $\eta_{\text{logit}} \mathbb{E}_x \|a(x)\mathbf{s}\|_{F_x}^2 / E_0$.

Remark D.5 (Shared-direction consistency under homogeneous metrics)**.** *Equation (56) controls the pulled-back residual, not variation of $B_x \mathbf{s}$. For example, replacing J_x by $-J_x$ preserves G_x but reverses the pullback, so equal metrics need not yield a common correction direction. For a consistent nonzero $\mathbf{w}_0$, normalize $\mathbf{w} = \mathbf{w}_0 / \|\mathbf{w}_0\|_G$ and $a_h(x) = a(x)\|\mathbf{w}_0\|_G$. To use this component in Assumption D.5, one must additionally verify $Q\mathbf{w} = \mathbf{w}$, zero residual mean, and weighted orthogonality. The bound on $\mathbf{e}_h(x)$ alone does not supply these moment conditions.*

From reward structure to recovery. The argument gives a sequence of sufficient conditions: an exact regularized optimum gives a next-token tilt relation; a shared success profile and a small relative residual give a shared logit component; a consistent pullback gives a shared representation component. The final lower bound separately requires off-principal localization and the moment conditions of Assumptions D.4–D.5. Each condition concerns a different object and must be checked independently when applying the analysis.

Connection to practical RL teachers. The tilt identities concern the exact optimum with $\beta > 0$. For trained checkpoints, including DAPO checkpoints with zero explicit KL coefficient, the fitting and recovery results take $\mathbf{t}(x)$ as given and apply under their stated local and structural conditions. Clipping and small updates motivate inspecting a local regime; they do not imply an exact tilt identity, metric homogeneity, or a shared direction for a trained checkpoint.

D.4 Single-vector recovery: an identity and a lower bound

The quadratic optimum is generally metric-weighted. Under exact homogeneity, its explained target energy has the signal-to-fluctuation form in Section D.4.1 and the sufficient geometric bound in Section D.4.2.

Proposition D.2 (Optimal steering vector)**.** *Assume finite target energy and finite C_{irr}. If $T^\top \bar{G} T$ is positive definite, where $\bar{G} = \mathbb{E}_x[G_x]$, the unique minimizer of the quadratic surrogate over $\mathcal{E}$ is $\boldsymbol{\delta}^\star = T(T^\top \bar{G} T)^{-1} T^\top \mathbb{E}_x[G_x \boldsymbol{\Delta}(x)]$. For $T = I_d$, it is $\bar{G}^{-1} \mathbb{E}_x[G_x \boldsymbol{\Delta}(x)]$. Under exact homogeneity,*

$$\boldsymbol{\delta}^\star = \boldsymbol{\mu} := \mathbb{E}_{x \sim \mathcal{D}}[\boldsymbol{\Delta}(x)].$$

This optimizes the quadratic surrogate without a trust-region constraint, not necessarily the original nonlinear KL objective.

Proof. Writing $\boldsymbol{\delta} = Tz$, the normal equation is $T^\top \bar{G} T z = T^\top \mathbb{E}_x[G_x \boldsymbol{\Delta}(x)]$. Positive definiteness gives the unique solution. If $G_x = G$ on $\mathcal{E}$, $\mathbb{E}_x[\boldsymbol{\Delta}(x)] \in \mathcal{E}$ solves the equation. □

Definition D.2 (Recovery ratio)**.** *In the constant-metric quadratic model, let* $E_{\mathrm{tar}} = \mathbb{E}_x\|\boldsymbol{\Delta}(x)\|_G^2 \in (0, \infty)$ *and define*

$$\rho := 1 - \frac{\mathbb{E}_x\|\boldsymbol{\Delta}(x) - \boldsymbol{\delta}^\star\|_G^2}{E_{\mathrm{tar}}} \in [0, 1].$$

This is the fraction of locally representable target energy explained by the optimal constant vector, excluding C_{irr}*. In particular,*

$$\frac{\mathcal{L}_{\mathrm{quad}}(0) - \mathcal{L}_{\mathrm{quad}}(\boldsymbol{\delta}^\star)}{\mathcal{L}_{\mathrm{quad}}(0)} = \frac{\rho E_{\mathrm{tar}}}{E_{\mathrm{tar}} + 2C_{\mathrm{irr}}}.$$

Thus, even perfect target-energy recovery need not remove irreducible logit error. Exact teacher reproduction additionally requires local reachability and negligible approximation error. Neither ρ *nor the displayed surrogate-loss ratio equals an empirical task-score gain recovery rate without further assumptions connecting distributions to task performance. Remark D.1(i) gives the varying-metric counterpart.*

Locally admissible interventions. Let $\mathcal{T} \subseteq \mathcal{E}$ be a specified trust region containing zero. Its best constant-metric target-energy recovery is

$$\rho_{\mathcal{T}} := 1 - \frac{\inf_{\boldsymbol{\delta} \in \mathcal{T}} \mathbb{E}_x\|\boldsymbol{\delta} - \boldsymbol{\Delta}(x)\|_G^2}{E_{\mathrm{tar}}}, \qquad 0 \le \rho_{\mathcal{T}} \le \rho.$$

The inequalities follow because zero is feasible and $\mathcal{T} \subseteq \mathcal{E}$. If $\boldsymbol{\mu} \in \mathcal{T}$, the two recovery values coincide. Otherwise, unconstrained surrogate recovery is only an upper bound on what this trust region permits. Relating either value to nonlinear KL additionally requires the remainder bounds to hold at the evaluated interventions.

D.4.1 Signal-to-fluctuation form

Theorem D.1 (Recovery as a signal-to-fluctuation ratio)**.** *In the constant-metric quadratic model, let* $\tilde{\boldsymbol{\Delta}}(x) = \boldsymbol{\Delta}(x) - \boldsymbol{\mu}$*, with* $\mathbb{E}_x[\tilde{\boldsymbol{\Delta}}(x)] = \mathbf{0}$ *and positive finite total target energy. Then*

$$\boxed{\begin{aligned} \rho &= \frac{\|\boldsymbol{\mu}\|_G^2}{\|\boldsymbol{\mu}\|_G^2 + \mathbb{E}_x\|\tilde{\boldsymbol{\Delta}}(x)\|_G^2} = \frac{\mathrm{SNR}}{1 + \mathrm{SNR}}, \\ \mathrm{SNR} &:= \frac{\|\boldsymbol{\mu}\|_G^2}{\mathbb{E}_x\|\tilde{\boldsymbol{\Delta}}(x)\|_G^2}. \end{aligned}} \tag{57}$$

If the fluctuation energy vanishes, set $\mathrm{SNR} = +\infty$ *and* $\rho = 1$*. For any target level* $u \in (0, 1)$*,* $\rho \ge u$ *if and only if* $\mathrm{SNR} \ge u/(1-u)$*, using this convention.*

Proof. Proposition D.2 gives $\boldsymbol{\delta}^\star = \boldsymbol{\mu}$. Since G is constant and the fluctuation has zero mean, $\mathbb{E}_x\|\boldsymbol{\Delta}(x)\|_G^2 = \|\boldsymbol{\mu}\|_G^2 + \mathbb{E}_x\|\tilde{\boldsymbol{\Delta}}(x)\|_G^2$. Substitution into Definition D.2 proves Equation (57). If the fluctuation energy is zero, positive total energy makes the mean energy positive. The threshold relation follows by rearranging $\rho = \mathrm{SNR}/(1 + \mathrm{SNR})$. □

Corollary D.2 (Spectral upper bound)**.** *In the same model, let* $\mathbf{u}(x) = G^{1/2}\boldsymbol{\Delta}(x)$*,* $\mathbf{m} = \mathbb{E}_x[\mathbf{u}(x)] = G^{1/2}\boldsymbol{\mu}$*, and* $\Sigma = \mathbb{E}_x[\mathbf{u}(x)\mathbf{u}(x)^\top]$*. For a weighted data operator or factor* M *with* $\Sigma = MM^\top$*,*

$$\rho = \frac{\|\mathbf{m}\|_2^2}{\operatorname{tr}\Sigma} \le \frac{\lambda_1(\Sigma)}{\operatorname{tr}\Sigma} = \frac{\sigma_1^2(M)}{\sum_i \sigma_i^2(M)}.$$

For $\mathbf{m} \neq 0$*, equality holds exactly when* $\mathbf{m}$ *lies in the leading eigenspace of* Σ *and* $\mathbf{m}^\top \operatorname{Cov}(\mathbf{u}(x))\mathbf{m} = 0$*. For* $\mathbf{m} = 0$ *and* $\operatorname{tr}\Sigma > 0$*, the inequality is strict.*

Proof. Since $\Sigma = \mathbf{m}\mathbf{m}^\top + \operatorname{Cov}(\mathbf{u}(x))$, one has $\mathbf{m}\mathbf{m}^\top \preceq \Sigma$. For $v = \mathbf{m}/\|\mathbf{m}\|_2$,

$$\lambda_1(\Sigma) \ge v^\top \Sigma v = \|\mathbf{m}\|_2^2 + v^\top \operatorname{Cov}(\mathbf{u}(x))v \ge \|\mathbf{m}\|_2^2.$$

Equality requires equality at both steps. When $\mathbf{m} = 0$, positive trace implies $\lambda_1(\Sigma) > 0$. Finally, $\operatorname{tr}\Sigma = E_{\mathrm{tar}}$ and the factorization gives the singular-value expression. □

Implications for capacity, gap, and depth. The leading spectral share is an upper bound, not generally the recovery of a constant vector. Collinear targets can have unit leading spectral share but low mean-energy recovery when their amplitudes vary or their signs cancel. Uniformly scaling a fixed target family by a nonzero scalar leaves ρ unchanged in a fixed metric. Thus, target magnitude or teacher–student gap alone does not impose a monotone capacity requirement. Injection depth can change the Jacobian, reachability, sensitivity metric, and target variation; the theorem does not assert a universal monotone relationship with depth. The observed gap and depth trends are empirical findings consistent with changes in these quantities, not direct consequences of the recovery identity.

D.4.2 Explicit three-constant bound

Assumptions D.4–D.5 specify a sufficient favorable regime through principal energy, complement residual energy, and amplitude consistency.

Theorem D.2 (Off-principal single-direction bound)**.** *In the constant-metric quadratic model, under Assumptions D.4–D.5 and their working-subspace compatibility conditions,*

$$\boxed{\rho \geq \frac{\kappa}{(1+\eta)(1+\varepsilon)}.} \tag{58}$$

Energy decomposition. Both moment conditions in Equation (37) are used. In particular, weighted orthogonality removes the mixed term in

$$\mathbb{E}_x\|a(x)\mathbf{w} + \boldsymbol{\xi}(x)\|_G^2 = \mathbb{E}_x[a(x)^2]\|\mathbf{w}\|_G^2 + \mathbb{E}_x\|\boldsymbol{\xi}(x)\|_G^2 + 2\mathbb{E}_x[a(x)\langle \mathbf{w}, \boldsymbol{\xi}(x)\rangle_G]. \tag{59}$$

Zero residual mean identifies $Q\boldsymbol{\mu} = \mathbb{E}_x[a(x)]\mathbf{w}$. Remark D.1 supplies a perturbation bound when the principal/-complement split is only approximately G-orthogonal.

Proof. By Theorem D.1, $\rho = \|\boldsymbol{\mu}\|_G^2/E_{\mathrm{tar}}$. The zero-mean residual gives $Q\boldsymbol{\mu} = \mathbb{E}_x[a(x)]\mathbf{w}$. Metric compatibility and $\|\mathbf{w}\|_G = 1$ imply

$$\|\boldsymbol{\mu}\|_G^2 \geq \|Q\boldsymbol{\mu}\|_G^2 = (\mathbb{E}_x[a(x)])^2 = \kappa\mathbb{E}_x[a(x)^2].$$

The energy split, localization bound, and weighted orthogonality give

$$E_{\mathrm{tar}} \leq (1+\varepsilon)\mathbb{E}_x\|Q\boldsymbol{\Delta}(x)\|_G^2 = (1+\varepsilon)(1+\eta)\mathbb{E}_x[a(x)^2].$$

The last factor is positive; division proves the result. □

What the bound establishes. Write $E_P = \mathbb{E}_x\|P\boldsymbol{\Delta}(x)\|_G^2$ and $E_Q = \mathbb{E}_x\|Q\boldsymbol{\Delta}(x)\|_G^2 > 0$. The proof gives complement recovery $\rho_Q = \|Q\boldsymbol{\mu}\|_G^2/E_Q = \kappa/(1+\eta)$. If $E_P > 0$, define $\rho_P = \|P\boldsymbol{\mu}\|_G^2/E_P$ and $e = E_P/E_Q$. Then the exact relation is

$$\rho = \frac{\rho_Q + e\rho_P}{1+e}.$$

If $E_P = 0$, $\rho = \rho_Q$. The theorem discards a nonnegative principal contribution and uses $e \leq \varepsilon$. Accordingly, low principal energy is a sufficient condition within this decomposition, not necessary for high constant-vector recovery: a constant nonzero target entirely in the principal subspace also has $\rho = 1$. The theorem quantifies a specified localization pattern; it does not derive that pattern from the RLVR objective or establish its prevalence.

Corollary D.3 (Coefficient-of-variation form)**.** *Under the same conditions,*

$$\rho \geq \frac{1}{(1+\mathrm{CV}(a)^2)(1+\eta)(1+\varepsilon)},$$

since $\kappa = 1/(1+\mathrm{CV}(a)^2)$.

Part II. Multi-Vector Steering: The Geometry of the Residual Subspace

Section D.5 extends constant-vector fitting to subspace constraints and distinguishes it from adaptive subspace approximation. Section D.6 relates principal structure to sensitivity and weight-induced activation changes.

D.5 Sequential-orthogonal steering and constrained fitting

The empirical protocol learns $\mathbf{v}^{(0)}, \ldots, \mathbf{v}^{(K-1)}$ sequentially. At each stage only the current multi-layer direction is injected; earlier directions define layer-wise Euclidean orthogonality constraints. Reprojection enforces these constraints after each optimizer step, and each direction is evaluated independently. The result below concerns exact fitting for a fixed target distribution and metric, not cumulative injection of the learned vectors.

Theorem D.3 (Constant-vector recovery under a subspace constraint)**.** *Let $\mathcal{V}_k \subseteq \mathcal{E}$ be a feasible subspace in the constant-metric quadratic model, and let R_k have full-column-rank columns spanning it. The exact constrained minimizer is*

$$\mathbf{v}^{(k)} = R_k(R_k^\top G R_k)^{-1} R_k^\top G\boldsymbol{\mu}.$$

For the single-site sequential protocol, $\mathcal{V}_k = \mathcal{E} \cap \mathrm{span}\{\mathbf{v}^{(j)} : j < k\}^{\perp_2}$. With $E_{\mathrm{tar}} > 0$, its standalone target-energy recovery satisfies

$$\rho^{(k)} = \frac{\|\mathbf{v}^{(k)}\|_G^2}{E_{\mathrm{tar}}}, \qquad \mathcal{V}_{k+1} \subseteq \mathcal{V}_k \implies \rho^{(k+1)} \le \rho^{(k)}. \tag{60}$$

If $G = cI$ on $\mathcal{E}$, $c > 0$, $\mathcal{V}_0 = \mathcal{E}$, and $\boldsymbol{\mu} \neq 0$, exact first-stage fitting gives $\mathbf{v}^{(0)} = \boldsymbol{\mu}$, and fitting in its Euclidean orthogonal complement gives $\mathbf{v}^{(1)} = \mathbf{0}$.

Proof. The parallel-axis identity splits the fitting objective into a constant fluctuation term and $\|\mathbf{v} - \boldsymbol{\mu}\|_G^2$. Writing $\mathbf{v} = R_k z$ gives $R_k^\top G(R_k z - \boldsymbol{\mu}) = 0$ and the stated solution. The residual is G-orthogonal to $\mathcal{V}_k$, hence $\langle \mathbf{v}^{(k)}, \boldsymbol{\mu}\rangle_G = \|\mathbf{v}^{(k)}\|_G^2$. Expanding the loss reduction yields Equation (60). Nesting the feasible subspaces cannot improve the optimum. In the isotropic case the second space is orthogonal to $\boldsymbol{\mu}$, whose projection is zero. If the feasible space has dimension zero, its unique minimizer is zero without applying the matrix-inverse formula. A zero vector introduces no further orthogonality constraint. □

For multiple sites, intersect the product of the layer-wise feasible spaces with the working subspace. The theorem applies to the stacked intervention and the full sensitivity metric, including cross-layer blocks.

Adaptive subspace approximation. For comparison, consider a shared subspace with unrestricted context-dependent coefficients. Let $1 \le K \le \dim \mathcal{E}$ and $\mathbf{u}(x) = G^{1/2}\boldsymbol{\Delta}(x)$, with $\Sigma = \mathbb{E}_x[\mathbf{u}(x)\mathbf{u}(x)^\top]$. Then

$$\min_{\substack{Q_K^\top Q_K = I_K \\ \mathrm{range}(Q_K) \subseteq G^{1/2}\mathcal{E}}} \mathbb{E}_x\|\mathbf{u}(x) - Q_K Q_K^\top \mathbf{u}(x)\|_2^2 = \mathrm{tr}\,\Sigma - \sum_{i=1}^{K} \lambda_i(\Sigma). \tag{61}$$

Indeed, the captured energy is $\mathrm{tr}(Q_K^\top \Sigma Q_K)$, maximized by a leading eigenspace. The fitted coefficients $Q_K^\top \mathbf{u}(x)$ depend on context. Thus, the optimal explained fraction is $\sum_{i=1}^{K} \lambda_i(\Sigma)/\,\mathrm{tr}\,\Sigma$ when the trace is positive. These are oracle coefficients for the quadratic target. A particular gated network must also represent and learn the required coefficient map, and layer-wise restrictions may further reduce its attainable fit. The spectral expression is therefore a characterization of a more flexible approximation class, not a guarantee for a trained gating architecture.

Interpretation of later steering directions. Nonzero standalone performance of later Euclidean-orthogonal vectors establishes the existence of alternative useful interventions. It does not establish that they recover distinct spectral components of the teacher target. Even a deterministic target can admit more than one useful Euclidean-orthogonal intervention under an anisotropic metric. For example, with $\mathcal{E} = \mathbb{R}^2$, $\boldsymbol{\Delta}(x) = (1, 0)^\top$, and $G = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, Theorem D.3 gives $\mathbf{v}^{(0)} = (1, 0)^\top$ and $\mathbf{v}^{(1)} = (0, 1)^\top$, with standalone recovery 1 and $1/2$, respectively, although the target second moment has rank one. In the empirical setting, anisotropy, nonlinear responses, evolving sampled contexts, and approximate optimization can all affect the learned directions; task scores are also distinct from quadratic energy recovery. The observed decline with extraction order and its model-scale dependence are consequently descriptions of independently effective interventions, not estimates of target spectral rank or intrinsic manifold dimension.

D.6 Effective directions and the principal subspace

The main-text steering measurements place approximately 1%–2% of the learned vectors' Euclidean energy in a top-10% activation principal subspace, below the approximately 10% isotropic reference. This observation concerns the learned vectors. The recovery bound instead uses sensitivity-weighted energy of the context-dependent target family. We now state the metric comparison needed to connect such quantities when the shift objects are matched.

Definition D.3 (Variance and sensitivity metrics)**.** *For a specified base-model reference context distribution $\mathcal{D}_{\mathrm{ref}}$, let $\bar{\mathbf{h}}_\ell = \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{ref}}}[\mathbf{h}_\ell(x)]$ and define*

$$C := \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{ref}}}[(\mathbf{h}_\ell(x) - \bar{\mathbf{h}}_\ell)(\mathbf{h}_\ell(x) - \bar{\mathbf{h}}_\ell)^\top], \qquad G := \mathbb{E}_{x\sim\mathcal{D}}[G_x].$$

The leading eigenspace of C defines P, with $Q = I - P$; fix a choice of basis if the cutoff eigenvalue is degenerate. The matrix G measures output sensitivity, and at the unembedding input equals $U^\top \mathbb{E}_x[F_x]U$. The reference distribution used to estimate covariance and the distribution used for fitting must be specified; they need not be identical. An uncentered implementation uses $C + \bar{\mathbf{h}}_\ell \bar{\mathbf{h}}_\ell^\top$ instead and must report its basis accordingly. High activation variance is not equivalent to high or low output sensitivity.

Proposition D.3 (Sensitivity-weighted principal-energy bound)**.** *On the working subspace, define*

$$g_P = \sup_{v \in \mathcal{E} \cap \mathrm{range}(P) \setminus \{0\}} \frac{v^\top G v}{\|v\|_2^2}, \qquad g_Q = \inf_{v \in \mathrm{span}\{Q\mathbf{\Delta}(x)\} \setminus \{0\}} \frac{v^\top G v}{\|v\|_2^2} > 0,$$

with $g_P = 0$ if $P\mathcal{E} = \{0\}$. If $\mathbb{E}_x \|Q\mathbf{\Delta}(x)\|_2^2 > 0$, then

$$e_G := \frac{\mathbb{E}_x \|P\mathbf{\Delta}(x)\|_G^2}{\mathbb{E}_x \|Q\mathbf{\Delta}(x)\|_G^2} \le \frac{g_P}{g_Q} \frac{\mathbb{E}_x \|P\mathbf{\Delta}(x)\|_2^2}{\mathbb{E}_x \|Q\mathbf{\Delta}(x)\|_2^2}.$$

If the right-hand side is small, it supplies a small admissible $\varepsilon \ge e_G$ for the localization part of Assumption D.4. Its metric-orthogonality condition remains separate.

Proof. At each context, $\|P\mathbf{\Delta}(x)\|_G^2 \le g_P \|P\mathbf{\Delta}(x)\|_2^2$ and $\|Q\mathbf{\Delta}(x)\|_G^2 \ge g_Q \|Q\mathbf{\Delta}(x)\|_2^2$. Averaging and dividing by the positive denominator proves the claim; the zero-principal-space case is immediate. □

Sensitivity and localization are distinct. If $J_x v = c(x)\mathbf{1}_V + e(x)$, softmax gauge invariance gives $v^\top G_x v = \|e(x)\|_{F_x}^2$. A small g_P requires an appropriate uniform sensitivity bound over the principal subspace, not merely this identity for one vector. The ratio g_P/g_Q can otherwise be large, so a small Euclidean PC fraction need not imply small sensitivity-weighted intrusion. Nor does localization of a learned mean imply localization of its target family. With $G = I$, $P = e_1 e_1^\top$, and equiprobable targets $\mathbf{\Delta}_\pm = (\pm c, 1)^\top$, the optimal constant vector is $(0, 1)^\top$ and has zero principal energy, whereas target principal-to-complement energy is c^2 and recovery is $1/(1 + c^2)$. Measurements of learned-vector energy therefore do not alone verify Assumption D.4.

D.6.1 A WEIGHT-SPACE ACCOUNT OF ACTIVATION LEAKAGE

We relate a layer's weight displacement to its first-order activation change. This is distinct from both the fitted intervention and the pulled-back teacher target.

Weight-principal components. For $W \in \mathbb{R}^{d \times d'}$ with singular value decomposition $W = \sum_i s_i \mathbf{p}_i \mathbf{q}_i^\top$, define

$$P_W := \sum_{i \le k} \mathbf{p}_i \mathbf{p}_i^\top.$$

The split $\Delta W = P_W \Delta W + (I - P_W)\Delta W$ separates principal weight components from their complement. The bound $|s_i(W + \Delta W) - s_i(W)| \le \|\Delta W\|_2$ controls absolute spectral displacement but does not determine activation-principal localization.

Lemma D.6 (Weight components and first-order activation leakage)**.** *Consider $\mathbf{h}_\ell(x) = \phi(W\mathbf{c}(x))$ with element-wise differentiable ϕ, holding $\mathbf{c}(x)$ fixed while perturbing W. Let $D(x) = \mathrm{diag}(\phi'(W\mathbf{c}(x)))$ have bounded norm at the evaluated points. For the first-order shift $\mathbf{\Delta}_\mathbf{h}(x) = D(x)\Delta W \mathbf{c}(x)$,*

$$\|P\mathbf{\Delta}_\mathbf{h}(x)\|_2 \le \underbrace{\|D(x)\|_2 \|P_W \Delta W\|_2 \|\mathbf{c}(x)\|_2}_{\text{on-principal weight mass}} + \underbrace{\|PD(x)(I - P_W)\|_2 \|\Delta W\|_2 \|\mathbf{c}(x)\|_2}_{\text{cross-subspace leakage}}. \tag{62}$$

If $\|\mathbf{\Delta}_\mathbf{h}(x)\|_2 > 0$ and the right-hand side is at most $\xi \|\mathbf{\Delta}_\mathbf{h}(x)\|_2$, then $\|P\mathbf{\Delta}_\mathbf{h}(x)\|_2 / \|\mathbf{\Delta}_\mathbf{h}(x)\|_2 \le \xi$.

Proof. Decompose

$$P\mathbf{\Delta}_\mathbf{h}(x) = PD(x)P_W \Delta W \mathbf{c}(x) + PD(x)(I - P_W)\Delta W \mathbf{c}(x).$$

Apply the triangle inequality, submultiplicativity, and $\|P\|_2 \le 1$. The relative bound follows by division when the total shift is nonzero. □

The leakage factor separates as $PD(x)(I - P_W) = (P - P_W)D(x)(I - P_W) + P_W D(x)(I - P_W)$, exposing basis mismatch and Jacobian mixing. Small weight-update norm or zero principal weight mass alone does not make this factor small relative to the total activation change. Extending the result beyond first order requires a nonlinear remainder bound; changing earlier layers additionally perturbs $\mathbf{c}(x)$. The lemma is a conditional first-order decomposition, not a deduction of activation localization from small weight updates.

PART III. GEOMETRIC ANALYSIS OF ALPHA-STABLER

Parts I–II analyze frozen-model intervention recovery. We now study a distinct parameter-training procedure: a fixed-reference activation diagnostic and local backward gradient projection. The diagnostic's energy identity and the projection's local properties are exact; the link to parameter-induced activation changes is conditional, and overall warning and stabilization performance is empirical.

D.7 PRINCIPAL-SUBSPACE MONITORING AND GRADIENT CONTROL

Warm-up estimation and fixed reference geometry. During the first $T_{\text{warm}} = 50$ parameter updates, the actor trains without projection. Its generated sequences are also evaluated by the frozen initial base model to estimate reference token-activation covariance. After warm-up, let $\mathcal{U}_\ell \in \mathbb{R}^{d\times r}$ contain the selected leading orthonormal eigenvectors, with $r = \lceil 0.10d \rceil$. This is the main-text basis $U_{\ell,r}$; the retained fraction refers to hidden dimensions, not explained variance. Set $P_\ell = \mathcal{U}_\ell\mathcal{U}_\ell^\top$ and $Q_\ell = I - P_\ell$ and hold them fixed thereafter. The analysis conditions on the estimated projectors. It neither assumes they equal population projectors nor supplies a finite-sample PCA guarantee. During each controlled backward pass and optimizer step, both projectors and flags are fixed, and their estimation and selection are detached from differentiation.

Token-level principal-subspace intrusion. For identical actor/base token sequences, stack the analyzed valid token-level activation differences as rows of $X_\ell^{(t)} \in \mathbb{R}^{N\times d}$. Consistent with Equation (6),

$$\text{PSI}_\ell^{(t)} = \frac{\|X_\ell^{(t)}\mathcal{U}_\ell\|_F^2}{\|X_\ell^{(t)}\|_F^2 + \epsilon_{\text{num}}}, \qquad \epsilon_{\text{num}} > 0. \tag{63}$$

The stabilizer is the main-text numerical ϵ, not the theoretical intrusion bound ε. Energies are summed over tokens before normalization; averaging vectors first would permit cancellation. Aggregate PSI sums principal and total shift energies across monitored layers before taking their ratio. The identities below hold for the analyzed shift collection. If monitoring uses a token subset, they are exact for that subset, not automatically for all tokens or the population. The sampling and distributed aggregation protocol must be specified with the implementation.

Corollary D.4 (Energy interpretation and metric comparison)**.** *Fix a layer. Let $A = \|XP\|_F^2$, $B = \|XQ\|_F^2$, and $A + B > 0$, and set $\psi_0 = A/(A + B)$. For $B > 0$, define $e_E = A/B$. Then*

$$\psi_0 = \frac{e_E}{1 + e_E}, \qquad 0 \le \psi_0 - \text{PSI} = \psi_0\frac{\epsilon_{\text{num}}}{A + B + \epsilon_{\text{num}}}. \tag{64}$$

Write each row of X as $\mathbf{\Delta}_i^\top$. For the same shifts with $B > 0$, define $e_G = \sum_i \|P\mathbf{\Delta}_i\|_G^2 / \sum_i \|Q\mathbf{\Delta}_i\|_G^2$. Suppose $0 < m_G \le M_G$ bound the metric on the span of the projected shifts: $m_G\|v\|_2^2 \le \|v\|_G^2 \le M_G\|v\|_2^2$. With $\chi = M_G/m_G$,

$$\frac{e_G}{\chi} \le e_E \le \chi e_G. \tag{65}$$

If this same shift collection satisfies $e_G \le \varepsilon$, then $\text{PSI} \le \psi_0 \le \chi\varepsilon/(1 + \chi\varepsilon)$.

Proof. Orthogonality gives $\|X\|_F^2 = A + B$ and $\|X\mathcal{U}\|_F^2 = A$, proving the energy identities by substitution. The metric bounds give $m_G A \le \sum_i \|P\mathbf{\Delta}_i\|_G^2 \le M_G A$ and the corresponding bounds for B. Division proves Equation (65); the last claim follows from monotonicity of $u/(1 + u)$. If $B = 0$ and $A > 0$, $\psi_0 = 1$ and the finite-ratio comparison is inapplicable. If $A + B = 0$, the stabilized statistic is zero but contains no directional information. □

The shifts in this corollary are observed actor-to-base differences. To apply the bound from Parts I–II requires a justified relationship to $\mathbf{\Delta}(x)$, matched distributions, and appropriate metric bounds; the notation does not assert that relationship. A small PSI is a small directional energy fraction and need not imply small absolute activation displacement.

Predictor and threshold calibration. During an initially stable warm-up, retain detached actor-to-base shift matrices at checks every $M = 3$ updates, subject to $N > 0$ and $\|X_\ell^{(t)}\|_F^2/N > \eta_{\min}$. The latter is a floor on the average squared token-level shift norm, not the energy of an averaged shift. After the reference bases are frozen, evaluate the retained matrices using Equation (63); let $\mathcal{W}_\ell$ be the valid warm-up PSI values, as in Algorithm 1. Set

$$\begin{aligned} m_\ell &= \text{median}_{\psi\in\mathcal{W}_\ell}\,\psi, \qquad & s_\ell &= 1.4826\,\text{median}_{\psi\in\mathcal{W}_\ell}\,|\psi - m_\ell| + \epsilon_{\text{num}}, \\ \tau_\ell^{\text{on}} &= m_\ell + 3s_\ell, & \tau_\ell^{\text{safe}} &= m_\ell + 2s_\ell. \end{aligned} \tag{66}$$

Calibration is invalid if any $\mathcal{W}_\ell$ is empty or $0 < \tau_\ell^{\text{safe}} < \tau_\ell^{\text{on}} < 1$ fails; training stops for recalibration rather than using such thresholds. After calibration, bases and thresholds are fixed and retained records can be released.

Every M updates, valid observations update $\bar\psi_\ell \leftarrow 0.95\bar\psi_\ell + 0.05\text{PSI}_\ell$, initialized at m_ℓ. The flag $a_\ell \in \{0, 1\}$ starts at zero. It activates after $p = 3$ consecutive valid checks with $\bar\psi_\ell > \tau_\ell^{\text{on}}$ and releases when a valid check gives $\bar\psi_\ell < \tau_\ell^{\text{safe}}$. A valid check with $\bar\psi_\ell \le \tau_\ell^{\text{on}}$, or an invalid check, resets the consecutive-warning count; an invalid check leaves the moving average and flag unchanged. Flags are held between checks and fixed throughout the associated gradient accumulation, backward pass, and optimizer update. These rules define a detector with smoothing, persistence, and hysteresis. The MAD multiplier does not provide a false-alarm probability for dependent training observations. An initially stable calibration window is an operational requirement, and useful lead times or error rates require empirical evaluation rather than following from Equation (66).

Controller: backward-only gradient projection. Let $G_\ell^{(t)} \in \mathbb{R}^{N\times d}$ be the incoming activation gradient at a monitored output before its local control operation. This is distinct from the sensitivity metric G in Parts I–II. When several sites are controlled, the incoming gradient already includes any downstream modifications. With a fixed basis and detached flag, pass upstream

$$\widetilde{G}_\ell^{(t)} = G_\ell^{(t)} - a_\ell^{(t)}(G_\ell^{(t)}\mathcal{U}_\ell)\mathcal{U}_\ell^\top. \tag{67}$$

This hook leaves the forward mapping unchanged at fixed model parameters. At an active site it applies $G_\ell^{(t)}Q_\ell$; at an inactive site it leaves the incoming gradient unchanged. Future forward activations can change through the resulting parameter updates.

Proposition D.4 (Minimal-change removal of principal gradients)**.** *For a Euclidean orthogonal projector P and $Q = I - P$, the unique solution of*

$$\underset{Z:\, ZP=0}{\text{minimize}}\ \frac{1}{2}\|Z - G_\ell\|_F^2 \tag{68}$$

is $\widetilde{G}_\ell = G_\ell Q$, and

$$\widetilde{G}_\ell P = 0, \qquad \widetilde{G}_\ell Q = G_\ell Q, \qquad \|\widetilde{G}_\ell\|_F^2 = \|G_\ell\|_F^2 - \|G_\ell P\|_F^2. \tag{69}$$

Proof. Feasibility gives $Z = ZQ$. Principal and complement components are Frobenius-orthogonal, hence $\|Z - G_\ell\|_F^2 = \|Z - G_\ell Q\|_F^2 + \|G_\ell P\|_F^2$. The unique minimizer is $Z = G_\ell Q$. The remaining identities follow from $P^2 = P$, $Q^2 = Q$, and $PQ = 0$. □

The proposition concerns the incoming gradient at one hook. It preserves that gradient's complement and cannot increase its local norm. It does not imply a smaller parameter-gradient norm, preservation of the uncontrolled network's complete gradient, or an analogous projection of future activation changes. Those quantities also depend on network Jacobians, other control sites, and the optimizer.

From projected gradients to activation changes. We analyze one post-warm-up SGD step on a fixed batch. This provides a local relation between a projected activation gradient and the forward displacement it induces, rather than an optimizer-independent stability guarantee.

Lemma D.7 (Local principal leakage under projected SGD)**.** *Fix a token batch, one active control site, and a parameter block ϑ. Suppose the loss depends on this block only through the selected activations, all other parameters are held fixed, and $h(\vartheta)$ is twice continuously differentiable locally. Stack token activations as $h = \mathrm{vec}(H^\top)$, and define $J = \partial h/\partial\vartheta$, $\Pi = I_N \otimes P$, $\mathcal{Q} = I - \Pi$, and $g = \nabla_h \mathcal{L}$. For a plain SGD step with $\alpha \geq 0$ whose update segment stays within a neighborhood of bounded second derivative,*

$$\delta\vartheta = -\alpha J^\top \mathcal{Q} g, \qquad \delta h = -\alpha K \mathcal{Q} g + R, \qquad K = JJ^\top, \tag{70}$$

where $\|R\|_2 \leq c_h\alpha^2\|J^\top\mathcal{Q}g\|_2^2$ for a local constant c_h. Consequently,

$$\|\Pi\delta h\|_2 \leq \alpha\|\Pi K\mathcal{Q}\|_2\|\mathcal{Q}g\|_2 + c_h\alpha^2\|J^\top\mathcal{Q}g\|_2^2. \tag{71}$$

Proof. The single-path dependency gives the modified parameter gradient $J^\top\mathcal{Q}g$. Taylor's theorem for the vector-valued map gives $h(\vartheta + \delta\vartheta) - h(\vartheta) = J\delta\vartheta + R$, with $\|R\|_2 \leq c_h\|\delta\vartheta\|_2^2$; one may take c_h as half a uniform bilinear operator-norm bound on its second derivative along the segment. Substitute the SGD step, apply Π, and use $\|\Pi\|_2 \leq 1$ and submultiplicativity to obtain Equation (71). □

What projection does and does not preserve. First-order principal displacement is controlled by $\Pi K\mathcal{Q}$, not by $\Pi\mathcal{Q} = 0$ alone. Relative to an unprojected SGD step at the same state, batch, and gradient,

$$\delta h_{\mathrm{proj}} - \delta h_{\mathrm{plain}} = \alpha K\Pi g + (R_{\mathrm{proj}} - R_{\mathrm{plain}}).$$

Thus the first-order difference in complement displacement is $\alpha\mathcal{Q}K\Pi g$, which need not vanish. If $\Pi K\mathcal{Q} = 0$, symmetry of K also gives $\mathcal{Q}K\Pi = 0$: the projected update then has no first-order principal displacement and preserves the plain step's complement displacement to first order. Without this additional decoupling, the lemma is a bound, not a guarantee of either preservation property or a comparison showing less leakage than the unprojected step.

Preconditioning and additional update terms. For a fixed realized parameter-space preconditioner D and an additional direction b, write a particular update as $\delta\vartheta = -\alpha(DJ^\top\mathcal{Q}g + b)$. The same Taylor argument gives, provided its update segment satisfies the same local regularity condition,

$$\begin{aligned}\|\Pi\delta h\|_2 \leq{}& \alpha\|\Pi JDJ^\top\mathcal{Q}\|_2\|\mathcal{Q}g\|_2 + \alpha\|\Pi Jb\|_2 \\ &+ c_h\alpha^2\|DJ^\top\mathcal{Q}g + b\|_2^2.\end{aligned}$$

The first-order coupling is now $JDJ^\top$; b may contain optimizer-state or other update contributions not represented by the current projected gradient. This is a conditional estimate for a specified step, not a bound on how an adaptive preconditioner or its state evolves. Applying it to a concrete optimizer requires its actual update decomposition. Multiple control sites likewise require the full modified gradient flow; the single-site SGD lemma does not supply a joint guarantee for the complete training algorithm.

Local leakage versus cumulative PSI and stability. The lemma bounds one update on a fixed batch, whereas PSI measures actor-to-base differences accumulated over training and is a ratio of energies. Even zero new principal displacement need not decrease PSI: for a one-token difference $(1, 1)$ with the first coordinate principal, a complement-only change $(0, -\alpha)$, $0 < \alpha < 1$, raises the unstabilized PSI from $1/2$ to $1/(1 + (1 - \alpha)^2)$. Long-run guarantees would additionally require uniform control of couplings, update magnitudes, remainders, optimizer state, and changing observation distributions. We therefore treat warning utility and training stabilization as empirical outcomes in Figures 7(c) and 8, not consequences of a monotone-PSI or no-collapse theorem.

Implementation and overhead. The mathematical properties require aligned actor/base observations and flags fixed before the controlled backward pass; they do not require a particular graph-retention strategy. Reusing a compatible actor forward and using a separate detached monitoring pass have different runtime and memory costs. Neither the projection identity nor the absence of additional trainable parameters bounds end-to-end overhead. Such claims require measurements that include reference evaluation, any additional actor pass, covariance estimation, calibration storage, and communication.

D.8 Summary of theoretical and empirical results

Table 5 records the object and scope of each conclusion. Parts I–II characterize conditional target-energy recovery and alternative intervention classes. Part III establishes finite-collection energy identities, exact local gradient-projection properties, and a conditional single-step relationship to activation displacement.

Summary. The analysis links three distinct questions while keeping their mathematical objects distinct. First, shared teacher targets can permit strong constant-vector recovery under local approximation, moment, and metric conditions. Second, constrained and adaptive fitting describe different intervention capacities; neither standalone orthogonality nor low learned-vector PC energy alone determines the teacher target's rank or localization. Third, Alpha-Stabler uses a fixed activation reference for monitoring and local gradient control. Its projection has exact minimal-change properties, while its effect on parameter-induced activation displacement depends on coupling and the optimizer. These conditional results organize the empirical capacity and localization observations and motivate the controller; they do not replace empirical evidence for task-score recovery, early warning, stability, or runtime overhead.

Table 5: Summary of recovery, steering geometry, and gradient-control results, with their assumptions and empirical scope.

Statement	**Reference**	**Scope / status**
Part I — fixed-vector recovery		
Distillation admits a weighted least-squares approximation	Proposition D.1	Fixed contexts; controlled network and KL remainders
The regularized optimum is an exponential reward tilt	Lemma D.2	Exact optimum of the stated objective, $\beta > 0$
Next-token ratios follow continuation normalizers	Lemma D.3	Exact under the tilt model
Shared profiles and consistent pullbacks give shared shifts	Lemmas D.4, D.5	Additional sufficient structural conditions
$\rho = \mathrm{SNR}/(1+\mathrm{SNR})$	Theorem D.1	Constant-metric surrogate target energy
Leading spectral energy share upper-bounds ρ	Corollary D.2	Upper bound, not a general equality
$\rho \geq \kappa/[(1+\eta)(1+\varepsilon)]$	Theorem D.2	Assumed localization, moments, and metric compatibility
Part II — geometry of steering interventions		
Constrained constant-vector fitting is a metric projection	Theorem D.3	Fixed target family and metric; exact optimization
Adaptive subspace recovery is a spectral partial sum	Equation (61)	Unrestricted context-dependent coefficients
Later orthogonal directions can remain useful	Section D.5	Empirical standalone interventions; not a rank estimate
Euclidean and sensitivity-weighted intrusion are related	Proposition D.3	Same targets; explicit metric constants
Weight components bound first-order activation leakage	Lemma D.6	Fixed input; basis-mismatch and mixing factors
Part III — monitoring and gradient control		
PSI measures Euclidean principal energy of analyzed shifts	Corollary D.4	Exact finite-collection identity; stabilizer accounted for
Projection minimally removes incoming principal gradients	Proposition D.4	Exact at each active hook with fixed projector
Principal activation displacement obeys a coupling bound	Lemma D.7	Single-site, fixed-batch, local plain SGD
Warm-up specifies a fixed reference and thresholds	Algorithm 1	Calibration procedure; no detection-error guarantee
PSI warns and control improves training stability	Figures 7(c) and 8	Empirical findings in the evaluated settings