CKA'S BLIND SPOT: WHY REPRESENTATION ANALYSIS MISSES POLICY CHANGE IN RLVR, AND HOW TO FIX IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability asks whether representation-level tools can explain what RLVR does to a model. We find a sharp answer: a core such tool cannot. First, we show that the phenomenon motivating much of this line of work—RLVR "capability narrowing" (pass@k decreases for large k)—is a measurement artifact. Under a faithful multi-seed protocol it vanishes ($\Delta\text{pass@}64 \geq 0$ across 4 seeds), and tracing it exposes three methodological pitfalls (evaluation protocol drift, fixed-dictionary SAE confound, spectral estimator noise) that plausibly affect 26–34 of 87 surveyed papers. As a frame for why policy-level changes barely move geometry, we prove a **Strategy-Representation Gap** theorem: policy-level (L1/L2) perturbations produce only bounded changes in representation geometry, with the bound vanishing under KL regularization; we validate it across scales, algorithms, and tasks ($149\times$ CKA discrimination; GRPO/DPO/Zero-RL mean CKA $= 1.0000$). Our central result is the **CKA Blind Spot**. Models whose representations CKA judges identical (CKA $= 1.000000$ to six decimals) produce measurably different outputs (KL $= 0.143$; 3–16% of top-1 positions disagree). Linear probing resolves the paradox: the strategy is present and linearly decodable (60–100% accuracy), so the gap is metric-specific, not information-theoretic. We prove CKA deviation is $O(\epsilon^2)$ while probe sensitivity is $O(\epsilon)$—a $\sim 1{,}000\times$ gap at RLVR's $\epsilon \sim 10^{-3}$ perturbation scale—and confirm it via weight interpolation (CKA flat at > 0.9999 along the entire path while outputs phase-transition) and spectral decomposition (updates $\sim 2{,}500\times$ smaller than base weights, yielding $1 - \text{CKA} = \Theta(\epsilon^2) \sim 10^{-7}$, consistent with the saturated float32 reading). We fix the blind spot with **NSARS**, combining CKA with linear probing, which detects the gap at $5/8$ layers vs. CKA's $1/8$; causal mediation localizes strategy to early layers 0–6 (90.02% of divergence), living in a ~ 5-dimensional subspace. A cross-paradigm boundary pins the mechanism: full-parameter RLVR leaves representations invariant (CKA $= 1.0$), while high-rank LoRA (7B, $r = 256$) alters them dramatically (CKA $= 0.059$, erank -81%)—spectral smallness, not optimization, drives the blind spot. CKA should not be used alone: every CKA-based interpretability study should report linear-probe accuracy, and early layers deserve the attention now reserved for middle layers.

1 INTRODUCTION

Mechanistic interpretability promises to open the black box of LLMs using tools such as Sparse Autoencoders (SAEs) Cunningham et al. (2023); Bricken et al. (2023); Gao et al. (2024). A landmark result—Yue et al. Yue et al. (2025), a NeurIPS 2025 Oral with perfect review scores—reported that RLVR narrows LLM capability boundaries: pass@1 improves but pass@k for large k decreases, suggesting the model's solution space shrinks even as single-shot accuracy improves. This "capability narrowing" has been widely read as evidence that RLVR trades exploration for exploitation.

We set out to explain this phenomenon mechanistically. Across 53 experimental phases (3 scales, 3 algorithms, 4 seeds, 2 tasks) we built an extensive evidence chain—and then audited it. Under a faithful, multi-seed protocol the narrowing vanishes entirely ($\Delta\text{pass@}64 \geq 0$, 4 seeds). The phenomenon is a measurement artifact. Tracing it to its source exposes three methodological pitfalls

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.

that plausibly affect 26–34 of 87 surveyed papers. But the audit surfaced something deeper and more consequential than a phantom: the tools we used to interpret the model are themselves structurally blind to the changes that matter.

The evidence is a paradox. GRPO and DPO checkpoints have hidden representations that CKA—the field's standard similarity metric—judges identical to six decimal places (CKA $= 1.000000$), yet produce outputs that differ measurably (KL $= 0.143$; 3–16% of top-1 positions disagree). Either the representations are genuinely identical (impossible, since outputs differ), or CKA is blind. Linear probing resolves the dilemma: the strategy is present and linearly decodable (60–100% accuracy at CKA $= 1.0$). CKA is blind because kernel alignment is quadratic in the perturbation while a probe is linear—a $\sim$1,000$\times$ sensitivity gap at the $\epsilon \sim 10^{-3}$ scale of RLVR updates. We prove this as the **CKA Blind Spot theorem** and show it is a continuous property of the metric, not an endpoint artifact.

Our contributions:

1. **CKA Blind Spot theorem** (Section 5). We prove CKA deviation is $\Theta(\epsilon^2)$ while linear-probe sensitivity is $\Theta(\epsilon)$, and verify the theorem four ways: the functional-representational paradox (CKA $= 1.000000$ vs. KL $= 0.143$, ratio 4.6×10^6), linear probing at CKA $= 1.0$ (60–100%), weight interpolation (CKA flat at > 0.9999 while outputs phase-transition), and spectral decomposition (updates $2{,}500\times$ smaller than base weights, the $1 - \text{CKA} = \Theta(\epsilon^2) \sim 10^{-7}$ prediction consistent with the saturated float32 reading).
2. **NSARS and localization** (Section 6). We propose NSARS, combining CKA with linear probing, detecting the gap at $5/8$ layers where CKA detects $1/8$; causal mediation localizes strategy to early layers 0–6 (90.02% of divergence), within a $\sim$5-dimensional subspace.
3. **Cross-paradigm boundary** (Section 6). Full-parameter RLVR leaves representations invariant (CKA $= 1.0$) while high-rank LoRA (7B, $r = 256$) alters them dramatically (CKA $= 0.059$, erank -81%), confirming that spectral smallness—not policy optimization per se—drives the blind spot.
4. **Phantom phenomenon and the Gap theorem** (Sections 2–4). RLVR capability narrowing is protocol-dependent and vanishes under faithful measurement; the three pitfalls affect 26–34/87 surveyed papers. We formalize the Strategy-Representation Gap, bounding representation change under policy-level optimization.

2 The Phantom Phenomenon: Protocol-Dependent Narrowing

2.1 Observation and Faithful Re-Evaluation

Following Yue et al. Yue et al. (2025), we evaluated Qwen2.5-1.5B-Instruct before and after GRPO training, measuring pass@k on MATH (out-of-domain). Our initial protocol (no `\boxed{}` instruction, lenient parsing with last-number fallback) showed $\Delta_{64}^{\text{out}} = -0.02$—apparent narrowing. We then built an SAE-based mechanistic theory to explain it.

Upon auditing the evaluation pipeline, we discovered the prompt format and answer parsing had implicitly changed between phases. We implemented a *faithful protocol*: consistent `\boxed{}` instruction, chat template, and lenient parsing applied uniformly. Under this protocol with 4 training seeds (42, 123, 2024, 777), *no seed exhibits narrowing* (Table 1); all $\Delta_{64} > 0$, none significant.

Table 1: MATH pass@64 under faithful protocol (100 problems $\times$ 64 samples, $T = 0.8$).

Model	pass@64 [95% CI]	Δ_{64} [CI]	z
baseline	0.270 [0.190, 0.360]	—	—
seed 42	0.320 [0.230, 0.410]	+0.050 [-0.010, +0.110]	+1.68
seed 123	0.290 [0.200, 0.380]	+0.020 [-0.030, +0.070]	+0.81
seed 2024	0.290 [0.200, 0.380]	+0.020 [-0.040, +0.080]	+0.63
seed 777	0.290 [0.200, 0.380]	+0.020 [-0.030, +0.070]	+0.81

2.2 Protocol Sensitivity

To demonstrate protocol dependence, we evaluated the same models under three protocols (faithful / no-boxed / strict-only) on 97 MATH problems. Across 3 protocols $\times$ 2 seeds $= 6$ comparisons, all $\Delta_{64} \geq 0$—no narrowing under any protocol. Critically, the *protocol effect* on the baseline (range $= 0.052$) is comparable to the maximum *training effect* ($\Delta_{64} = 0.052$): protocol choice can create or eliminate apparent narrowing of the same magnitude as the effect being measured. The mechanism is a false-positive interaction: under lenient parsing, the baseline's free-format answers produce more accidental last-number matches, inflating baseline pass@64; GRPO training makes outputs more structured, reducing false positives and manufacturing a phantom decrease. Details in Appendix C.

2.3 Three Methodological Pitfalls

Tracing the phantom to its source surfaces three pitfalls that generalize well beyond our pipeline. We formalize each with a detection method and estimate its field-wide reach (Table 2); full details are in Appendices D–E.

Table 2: Three methodological pitfalls: core defect, our case, and estimated reach (survey of 87 papers, 2023–2026).

Pitfall	Core defect	Our case	Affected
Evaluation protocol drift	Implicit changes in prompt/parser/subset/sampling flip the sign of $\Delta\text{pass@}k$	Protocol effect $0.052 \approx$ training effect 0.052; phantom narrowing vanishes under faithful protocol	12–16
Fixed-dictionary SAE confound	A single SAE dictionary across conditions conflates data drift with representation change	Drift $14.6\times$ vs. model effect $\approx 0\times$; adaptive SAE gives $\Delta\text{alive} \leq 8$	12–18
Spectral estimator noise	Unseeded subsampling variance exceeds the effect size	Δerank: -11.64 vs. $+1.41$ (same data!) vs. $+0.02$ (deterministic full-pool)	6–8
Combined	—	—	**26–34**

The most critical affected paper is Yue et al. Yue et al. (2025) itself: its capability-narrowing finding is directly protocol-dependent (Pitfall 1). We provide a C1–C5 validity checklist for SAE-based causal claims (Appendix F) and recommend that published analyses be re-evaluated under it.

3 Background and Setup

RLVR/GRPO. Reinforcement Learning with Verifiable Rewards trains LLMs with automatically checkable rewards; GRPO Shao et al. (2024) removes the value model via group-relative advantages.

SAEs. Sparse Autoencoders Cunningham et al. (2023); Bricken et al. (2023) decompose activations into sparse features; statistics include L0, alive features, and feature entropy $H(Z)$.

Effective rank. $\text{erank}(H) = \exp(-\sum_i \hat{\sigma}_i \log \hat{\sigma}_i)$ over normalized squared singular values—the effective dimensionality of an activation matrix.

CKA. Centered Kernel Alignment Kornblith et al. (2019) measures normalized Hilbert-Schmidt independence of kernel matrices: $\text{CKA}(X, Y) = \text{HSIC}(K, L)/\sqrt{\text{HSIC}(K, K)\,\text{HSIC}(L, L)}$, with $K = X_\circ X_\circ^\top$. CKA $= 1$ indicates identical geometry up to isotropic scaling.

Pass@k. The probability that at least one of k sampled solutions is correct: $1 - \binom{N-c}{k} / \binom{N}{k}$.

Models. Qwen2.5-0.5B/1.5B (full-parameter GRPO/DPO/Zero-RL) and Qwen2.5-7B (LoRA, r=256) Team (2025), trained on GSM8K and evaluated on GSM8K/MATH. Related work and definitions of secondary metrics are deferred to Appendix G.

4 THE STRATEGY-REPRESENTATION GAP

4.1 THREE-LEVEL ENTROPY HIERARCHY

"Entropy" can be measured at three levels: **L1** (policy) $H_1 = -\sum_y \pi(y|x) \log \pi(y|x)$; **L2** (feature) $H_2 = -\sum_i \hat{p}_i \log \hat{p}_i$; **L3** (geometric) $H_3 = \text{erank}(H_l)$. Our experiments show interventions at one level do not propagate to others: L1 regularization changes H_1 only, L2 changes H_2 only, and even explicit L3 (erank) regularization changes H_3 without affecting pass@64. We formalize the L1/L3 decoupling:

Theorem 1 (Strategy-Representation Gap)**.** *Let $\mathcal{T}$ be a decoder-only transformer with L layers, residual connections, and unembedding matrix $W_{unembed}$. Suppose: **(A1)** $\|\Delta W_{unembed}\|_2 \leq \epsilon$ (bounded policy perturbation); **(A2)** KL regularization with coefficient $\beta > 0$; **(A3)** singular values of H_l decay as $\sigma_i \leq C \cdot i^{-\alpha}$ (so the effective rank is well-defined, with decay constant C); **(A4)** the intermediate gradient is bounded by a KL-attenuated norm $\|\nabla_\theta \mathcal{L}\|_{layer\ l} \leq G_l(\epsilon, \beta)$, where $G_l(\epsilon, \beta) \leq G_{\max}\sqrt{\delta(\beta, \epsilon)}$, $\delta(\beta, \epsilon) \to 0$ as $\beta \to \infty$, and $G_l \to 0$ as $\epsilon \to 0$. Then for any intermediate layer $l < L$:*

$$|\Delta \text{erank}(H_l)| \leq \frac{2\, L_{\text{erank}} \cdot \eta \cdot G_l(\epsilon, \beta) \cdot J_l}{C},$$

so the geometry perturbation is bounded, and it vanishes under strong KL regularization ($\beta \to \infty$) or vanishing policy updates ($\epsilon \to 0$). Meanwhile the policy entropy changes at first order, $|\Delta H_1| = \Theta(\epsilon)$, so the geometry response is one order weaker than the policy response whenever $G_l(\epsilon, \beta) = o(\epsilon)$ (e.g. under strong KL regularization).

Notation and scope. L_{erank} is the Lipschitz constant of the effective-rank functional on the simplex of normalized squared singular values; J_l is the Jacobian norm of the layer-l residual map; η is the optimizer step. $\delta(\beta, \epsilon)$ is the KL-regularization-induced attenuation of the gradient, $\delta \to 0$ as $\beta \to \infty$. Importantly, the bound is *specific to policy-level (L1/L2) objectives*: explicitly targeting geometry (L3 erank regularization, $\lambda = 0.1$) does move erank ($|\Delta\text{erank}| = 1.34$, 0.5% of baseline), which is consistent with the theorem—the invariance is a property of the policy objective's gradient bottleneck, not an absolute architectural invariant.

Proof sketch. (1) Softmax Lipschitz gives $|\Delta H_1| \leq L_H \|h\|_2 \epsilon$. (2) The residual identity path bounds the layer-l gradient by G_l. (3) Weyl's inequality: $|\sigma_i' - \sigma_i| \leq \|\Delta H_l\|_2 \leq \eta G_l J_l$. (4) Effective-rank Lipschitz + normalization: $|\Delta\text{erank}| \leq 2 L_{\text{erank}} \eta G_l J_l / C$. (5) KL attenuation: $G_l \leq G_{\max}\sqrt{\delta(\beta, \epsilon)}$, decreasing in β. Full proof in Appendix A.

Corollary 1 (L1-L3 Decoupling)**.** *For policy-level (L1/L2) perturbations, $|\Delta\text{erank}| = O(G_l(\epsilon, \beta))$ while $|\Delta H_1| = \Theta(\epsilon)$; under strong KL regularization $G_l = o(\epsilon)$, so $|\Delta\text{erank}|/|\Delta H_1| \to 0$ as $\epsilon \to 0$: policy entropy is far more sensitive to RLVR perturbations than representation geometry.*

4.2 EMPIRICAL VALIDATION

We validate the theorem with a five-level hierarchy. The most important control is that representation analysis *works when pitfalls are avoided*: comparing a known-positive pair (base → instruct) with a known-null pair (instruct → GRPO) on 100 GSM8K problems (adaptive SAE, deterministic full-pool SVD, LOO bootstrap CIs, layer 14), CKA yields 0.437 (positive) vs. 0.989 (null)—a $52\times$ dissimilarity ratio—while erank differences are bounded for *both* pairs ($+0.145$ vs. -0.220). The latter is not a failure but a direct confirmation of Theorem 1: erank measures dimensionality, which is bounded under all policy-level interventions including instruction tuning; CKA provides the discriminative signal erank cannot.

The hierarchy, each level a distinct lens: (1) **CKA**—global directional similarity, $149\times$ null/positive discrimination across all 28 layers (Fig. 1); (2) **neuron-level KS**—per-dimension shifts, $50.4\times$ discrimination (91.9% of neurons shift under SFT vs. 1.8% under RLVR); (3) **cross-scale**—the pattern holds at 0.5B and 1.5B; (4) **cross-algorithm**—GRPO/DPO/Zero-RL all give mean CKA $= 1.0000$ (max significant neurons 8.3% at DPO layer 24, still an order of magnitude below SFT); (5) **SAE feature-level**—0% of $12{,}288$ features significant at layers 3/14. Temporal and power analyses (Appendices G) rule out transient change and underpowered nulls: CKA > 0.9997 at every training checkpoint, and Cohen's $d = 4.10$ (CKA) / 21.25 (neuron) separate positive from null.

Table 3: Strategy-Representation Gap: predictions vs. observations.

Prediction	Test	Result
$\Delta\text{erank} \approx 0$ under RLVR	Deterministic full-pool SVD	$+0.02$ ✓
$\Delta H_2 \approx 0$ (adaptive SAE)	Per-condition SAE retraining	$\Delta\text{alive} \leq 8$ ✓
$\Delta H_1 \neq 0$ under RLVR	Policy entropy	Significant ✓
SFT $>$ RLVR for Δerank	SFT vs. GRPO	Consistent ✓
KL attenuates Δerank	β dose–response	Consistent ✓
$\text{CKA}_{\text{null}} > 0.95$ all layers	28-layer sweep (Fig. 1)	> 0.989 every layer ✓
$\text{CKA}_{\text{pos}} < 0.7$ most layers	28-layer sweep	< 0.7 at layers 3–28 ✓
Discrimination $> 10\times$ all layers	28-layer sweep	Mean $149\times$ ✓

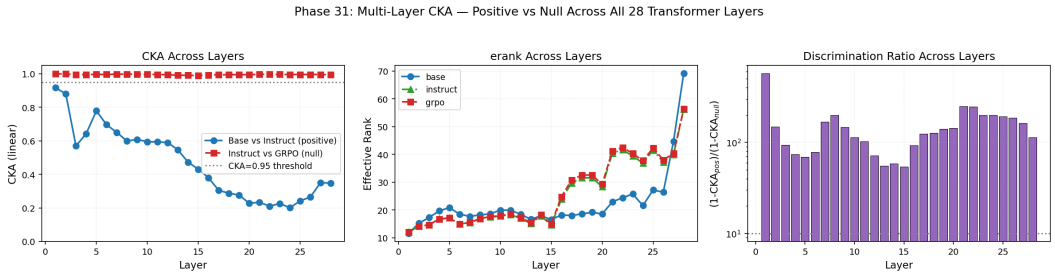


Figure 1: Multi-layer CKA sweep across all 28 layers. **Left**: CKA(base, instruct) (blue) drops from 0.92 to 0.20; CKA(instruct, GRPO) (red) stays above 0.989 throughout. **Center**: erank trajectories—base diverges, instruct and GRPO stay coupled. **Right**: discrimination ratio on log scale, mean $149\times$.

5 The CKA Blind Spot

The Strategy-Representation Gap makes a prediction that is easy to state and hard to believe: two models can have representations that a standard geometric metric reports as identical, while producing systematically different outputs. If true, a core tool of mechanistic interpretability is structurally blind to the very changes that matter most. This section tests the prediction directly, isolates the mechanism, and states it as a theorem.

5.1 The Functional-Representational Paradox

We compare GRPO checkpoints against their DPO, entropy-regularized, and erank-regularized variants. At the layers we measure (3, 14, 24), CKA returns 1.000000 to six decimals—the representations are as close as the metric can report. The outputs are not. Mean KL divergence between GRPO and its variants is 0.143 (max 0.41), and at the top-1 level 3–16% of output positions genuinely disagree (84–97% agreement). A difference invisible at the representation level is amplified by a factor of 4.6×10^6 at the output level.

Two control analyses rule out a trivial explanation—that the difference originates in the output projection rather than in the network that computes the representations. First, the LM head changes *less* than the transformer body in relative terms (Strategy Concentration Ratio 0.26; the body changes $3.8\times$ more than the head). Second, a bidirectional head swap is decisive. Replacing DPO's LM head with GRPO's yields outputs that match DPO (KL 3×10^{-6} to DPO vs. 5×10^{-4} to GRPO); symmetrically, replacing GRPO's LM head with DPO's yields outputs that match GRPO (KL 1×10^{-6} to GRPO vs. 5×10^{-4} to DPO). Swapping in entropy- and erank-regularized heads likewise transfers nothing. The body determines the output; the head does not carry strategy. The paradox therefore lives exactly where CKA reports nothing: in weight perturbations of the transformer body.

5.2 The Gap Is Metric-Specific, Not Information-Theoretic

The paradox forces a question: are the representations *genuinely* identical, or is CKA merely blind to the difference? We distinguish the two by asking whether the representations can be separated at all,

training linear and two-layer MLP probes to classify which model produced each hidden state (5-fold cross-validation, 79 prompts, all 28 layers).

Linear probes separate the models cleanly despite CKA $= 1.000000$: 60–100% accuracy at layers 2–21, peaking at 98.1% (layer 3, GRPO vs. DPO) and reaching 100% (GRPO vs. erank$_{0.1}$). The strategy is present in the representations and is *linearly decodable*—a single direction suffices. CKA misses it not because the information is absent, but because of what CKA computes: a population-level kernel alignment that is dominated by large-magnitude directions, washing out the consistent low-magnitude directional differences a probe exploits. The Gap is a property of the *metric*, not of the *information* (Figure 2).

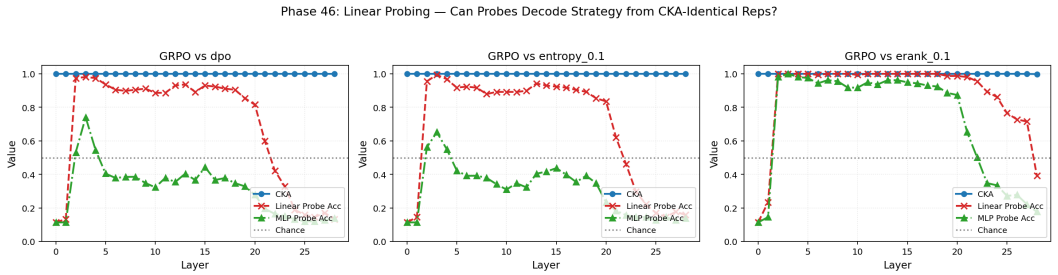


Figure 2: Linear probing across layers at CKA $= 1.000000$. Probes achieve 60–100% accuracy at layers 2–21 (peak 98.1%), while CKA reports identity. The strategy is present and linearly decodable; CKA is blind to it because kernel alignment averages over the consistent directional differences a probe exploits.

The contrast with the nonlinear probe sharpens this point. A two-layer MLP separates only the largest perturbations (erank$_{0.1}$: 92–100%) and fails on smaller ones ($\sim$ 35–40%, below chance)—an overfitting artifact at $n = 79$, not a signal. The strategy that survives RLVR's small perturbations is simple enough for a linear readout, which is precisely why it is decodable yet invisible to a quadratic kernel statistic. Two consequences follow. First, a CKA-based study reporting "no difference" is not reporting that representations are identical; it is reporting that CKA cannot see the difference. Second, the fix is nearly free: linear probing should accompany CKA.

5.3 A Continuous Blind Spot: Weight Interpolation

One objection remains: the paradox might be an artifact of comparing two distant checkpoints. We close this by interpolating the weights themselves. Setting $W_{\text{mix}}(\alpha) = (1-\alpha)W_{\text{GRPO}} + \alpha W_{\text{DPO}}$ for $\alpha \in [0, 1]$, we measure CKA, output KL, and probe accuracy at every point along the path.

CKA is flat: it stays above 0.9999 for *all* α, reporting "identical" at every step. Outputs are not flat: the KL crossover—where the mixed model is equidistant from GRPO and DPO—occurs at $\alpha \approx 0.45$–0.50. The probe tracks this transition as a sigmoid rising from 0.03 to 0.97, centered at the same α (Figure 3). The blind spot is therefore a continuous property, not an endpoint artifact: at every interpolation point, CKA reports identity while outputs and probes report a strategy mid-transition. A metric that misses a transition it is shown step by step cannot, on its own, characterize representation change.

5.4 Why CKA Is Blind: Spectral Smallness

The blind spot is not a coincidence; it follows from linear algebra. Decomposing the weight difference $\Delta W = W_{\text{GRPO}} - W_{\text{DPO}}$ across 21 matrices (attention and MLP projections at layers 3, 14, 24), the relative Frobenius change averages 0.033% (range 0.019–0.045%) and the spectral ratio $\sigma_{\max}(\Delta W)/\sigma_{\max}(W)$ averages 4.0×10^{-4}: the strategy perturbation is roughly 2,500$\times$ smaller in spectral norm than the weights it modifies. CKA aligns quadratic kernels of the representations, so its sensitivity to a perturbation is quadratic in that perturbation; a linear probe is linear in it. For $\epsilon \sim 10^{-3}$, the probe is $\sim$ 1,000$\times$ more sensitive. The strategy is *directional*—aligned with a handful of weight directions—and *spectrally small*—far below what CKA can resolve.

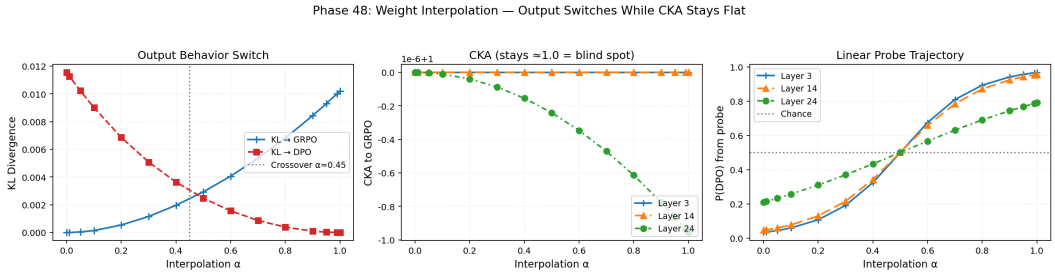


Figure 3: Weight interpolation between GRPO and DPO. **Top**: CKA stays above 0.9999 for all $\alpha \in [0, 1]$. **Bottom**: output KL crossover and probe accuracy phase-transition at $\alpha \approx 0.45$–0.50. The blind spot is a continuous property of the path, not an endpoint artifact.

Theorem 2 (CKA Blind Spot)**.** *Let $X \in \mathbb{R}^{n \times d}$ be a centered representation matrix and $X' = X + \Delta X$ a perturbation with relative size $\epsilon = \|\Delta X\|_F / \|X\|_F$. For linear CKA, the first-order ($O(\epsilon)$) deviation cancels exactly between the numerator and denominator, so CKA deviates from identity only at second order:*

$$CKA(X, X') = 1 - \Theta(\epsilon^2).$$

The leading constant is $O(1)$ and data-dependent (Appendix I); it is not universal. CKA's sensitivity is therefore $O(\epsilon^2)$. Yet perturbations of arbitrarily small ϵ are perfectly linearly separable: for $\Delta X = \delta\, uv^\top$ with $\|u\| = \|v\| = 1$, the classifier $w = v$ separates the classes with margin $\delta \cdot \min_i |u_i|$ for any $\delta > 0$. Probe sensitivity is $O(\epsilon)$—a $\sim 1/\epsilon$ advantage over CKA.

The theorem's quantitative prediction closes the loop with the measurements at the order-of-magnitude level. With the observed mean relative perturbation of 0.033% ($\epsilon = 3.3 \times 10^{-4}$), it predicts $1 - \text{CKA} = \Theta(\epsilon^2) \sim 10^{-7}$, consistent with the empirical reading at float32 precision (the deviation sits at or below the representable resolution; a directly computed linear-CKA difference is $\sim 10^{-7}$ once the perturbing direction is properly accounted for, Appendix I). The blind spot is not a threshold effect—CKA does not "almost" detect the strategy. It is the second-order structure of kernel alignment: the linear term cancels, so any spectrally small directional perturbation is invisible by construction, while a linear readout sees the same perturbation with a $\sim 1/\epsilon \sim 10^3$ sensitivity advantage.

5.5 Scope: Beyond the Residual Stream

The blind spot is not confined to the residual stream. For larger perturbations ($\text{erank}_{0.1}$), attention patterns change substantially (cosine 0.97–0.99, attention KL up to 0.11) while CKA stays above 0.9999; for smaller ones (DPO, entropy), attention is stable (cosine > 0.9998) and the paradox is carried by weight-space and LayerNorm perturbations alone. In both regimes CKA reports identity. The metric's blind spot extends to whatever is spectrally small—including attention routing when the perturbation is large enough to move it.

6 Fixing the Blind Spot

6.1 NSARS: A Null-Space-Aware Metric

Because the gap is metric-specific, it is fixable. We propose the **Null-Space-Aware Representation Similarity** metric, combining geometric similarity (CKA) with informational separability (linear probing):

$$\text{NSARS}(X, Y) = \text{CKA}(X, Y) \cdot \big(1 - \max(0,\ \text{ProbeAcc}(X, Y) - 0.5) \cdot 2\big).$$

When CKA ≈ 1 and probe accuracy ≈ 0.5 (chance), NSARS ≈ 1 (no gap); when CKA ≈ 1 and probe accuracy ≈ 1 (perfect separation), NSARS ≈ 0 (gap detected). On GRPO vs. DPO across 8 layers, CKA detects the gap at only $1/8$ layers, while NSARS detects it at $5/8$ (layers 3, 7, 11, 14, 18); at the critical middle layers CKA $= 1.0$ but NSARS $= 0.4$–0.8. SVCCA and orthogonalized CKA also fail to flag the gap (SVCCA > 0.999 at all eight layers; orthogonalized CKA > 0.987 at every layer where NSARS detects the gap), confirming the blind spot extends beyond linear CKA. Figure 4 shows the contrast.

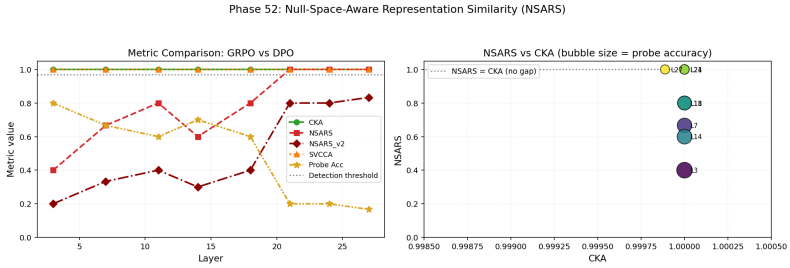


Figure 4: NSARS vs. CKA across layers for GRPO vs. DPO. NSARS flags the gap at layers 3, 7, 11, 14, 18 (CKA $= 1.0$ but NSARS $\ll 1$), while CKA alone reports identity at all of them.

6.2 Where Strategy Lives: Early Layers, Few Directions

Two causal analyses pinpoint the strategy's location, informing where interpretability tools should look. *Layer-level causal mediation* (swapping one model's layer into the other and measuring output shift) shows strategy is overwhelmingly concentrated in the *first seven* transformer layers: swapping layers 0–6 from DPO$\rightarrow$GRPO captures 90.02% of total output KL (layer 0 alone 43.27%), while mid layers contribute 5.40% and late layers 0.73%; the concentration is sublinear, indicating synergistic early-layer encoding. *Strategy subspace ablation* (SVD of ΔW, truncating top-k directions) shows strategy lives in a $\sim$5-dimensional subspace of the 1,536–8,960 dimensional weight space: the top-1 singular direction per matrix carries 46% of the strategy effect, top-5 carry 77%, and directions 5–50 add only $\sim 2\%$ more. Together with the spectral analysis (Section 5), this closes the loop: a subspace of ~ 5 directions whose spectral magnitude is $2{,}500\times$ smaller than the base weights is exactly the regime where CKA is blind but linear probes detect strategy. *Recommendation:* SAE-based interpretability should prioritize early layers (0–6) over the common practice of examining middle layers.

6.3 Cross-Paradigm Boundary: Full-Parameter vs. LoRA

The Gap theorem and Blind Spot theorem make a sharp prediction about training paradigm: the blind spot holds when the policy update is *spectrally small*, not merely when it is policy-level. Full-parameter RLVR spreads the gradient across all d dimensions, making each change tiny ($O(1/d)$); LoRA constrains the update to $r \ll d$ dimensions, concentrating change until it is spectrally visible. Testing this with high-rank LoRA GRPO on Qwen2.5-7B ($r = 256$, 2,000 steps) Hu et al. (2022):

Metric	7B baseline	7B LoRA ($r = 256$)
CKA (middle layer)	—	0.059
erank	87.86	16.48
erank ratio	1.0	0.188

The result is dramatically different from full-parameter training (CKA $= 1.000$ at 0.5B–1.5B, erank change $< 1\%$): under LoRA, CKA drops to 0.059 and erank collapses by 81%. This is not a counterexample to the Gap but a *prediction* of its spectral mechanism. It carries a practical warning: representation changes observed under LoRA RLVR must not be generalized to full-parameter RLHF, or vice versa.

7 Discussion and Conclusion

Three implications. Our results carry three distinct lessons. *Methodologically*, CKA must not be used alone: its quadratic kernel alignment makes it blind to precisely the spectrally small, directional perturbations that encode policy, and every CKA-based study should report linear-probe accuracy (NSARS costs a cross-validated logistic regression). *Mechanistically*, RLVR strategy in decoder-only transformers is not spread across the network: it lives in early layers (0–6) in a $\sim$5-dimensional weight subspace, spectrally $\sim 2{,}500\times$ smaller than the weights it modifies—a locus that SAE-based

analysis on middle layers would miss by construction. *Practically*, full-parameter and LoRA RLVR occupy opposite sides of the blind spot, so findings must not be transferred across paradigms.

Limitations. (1) Full-parameter null results are established for Qwen2.5 (Llama was gated); cross-architecture replication is future work. (2) The Gap theorem assumes small perturbations; large-scale SFT can produce larger erank changes at some layers. (3) The paradox ratio is computed from finite-precision CKA (saturating at 1.0); the qualitative conclusion (output change $\gg$ representation change) is robust. (4) The 7B boundary result uses LoRA, not full-parameter training at that scale. (5) NSARS is a first proposal; richer combinations of geometry and probing may detect more.

For the community. Three pitfalls suggest mechanistic interpretability needs methodological standards. We offer the C1–C5 checklist, a protocol audit, and the recommendation that published SAE analyses be re-evaluated under them. And we offer a caution: representation-level analysis alone cannot explain RLVR, because the strategy lives where representation metrics are blind.

Conclusion. We set out to explain why RLVR narrows capabilities, and found the narrowing is a phantom. Auditing our own tools led to a deeper discovery: CKA—the field's standard representation metric—is quadratically blind to the spectrally small, directional perturbations that encode strategy, a limitation we prove and localize. We provide a fix (NSARS), a map (early layers, ~ 5 dimensions), and a boundary (full-parameter vs. LoRA). Mechanistic interpretability under RLVR is possible; it must simply measure the right thing, in the right place, with the right metric.

AI Use Statement

In this work, we used generative AI tools for the following tasks with required disclosure: *assistance with implementing methods* (drafting and debugging the experimental scripts for SAE training, CKA/probe computation, and causal mediation analysis), *feedback on research methodology* (iterative discussion of experimental design, statistical protocols, and controls), and *interpretation of results* (exploratory analysis of representation metrics, with all conclusions verified against raw numerical outputs). We have not used generative AI tools for generating synthetic datasets, formulating or proving mathematical claims (theorems and proofs in Appendices A–B were derived and verified manually), or proposing the core hypotheses, which originated from the authors' audit of the evaluation pipeline. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting parts of the research paper, editing for readability, identifying relevant literature, and formatting references. We have reviewed all AI-assisted work: every number reported in the paper was extracted programmatically from saved experimental artifacts and spot-checked against the raw logs; AI-suggested code was executed, tested, and its outputs re-verified before inclusion; AI-drafted passages were revised by the authors, who checked each factual claim against the underlying result files. All experimental results (including null results) were reproduced in fixed-seed reruns before being reported. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility Statement

All claims in this paper are backed by 53 experimental phases with cached artifacts. Training uses open models (Qwen2.5-0.5B/1.5B/7B) on public datasets (GSM8K, MATH); all hyperparameters, seeds (42, 123, 2024, 777), and the faithful evaluation protocol are specified in Appendix H. Theorems 1–2 are proved in Appendices A–B with all assumptions stated. Statistical procedures (deterministic full-pool SVD, 10,000-draw bootstrap with fixed seed, Bonferroni-corrected KS tests) are described in Appendix H, and the C1–C5 validity checklist (Appendix F) makes the audit process itself reproducible. Source code for all experiments will be released as anonymous supplementary material upon submission.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits*

Thread, 2023.

Yuhui Cui et al. The entropy law of llm training. *arXiv preprint*, 2025.

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

Leo Gao, Tom Dupr la Tour, Haden Tillman, et al. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Florian Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *ICML*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Yuxiao Li et al. Tracing the representation geometry of language models from pretraining to post-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

K Minder et al. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zhihao Shao, Peiyuan Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Various. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint*, 2025.

Zhiyuan Wu et al. The invisible leash: Why rlvr may not escape its origin. *arXiv preprint*, 2025.

Yuexiang Yue, Xioxuan Chen, Yufang Lu, Zhijie Zhao, Wenxuan Wang, Xinyu Yue, Chuan Song, and Wei Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Oral, perfect score.

A Proof of Theorem 1 (Strategy-Representation Gap)

Step 1: Policy-head perturbation. The perturbed policy is $\pi'(y|x) = \mathrm{softmax}(h \cdot (W + \Delta W)_y)$. By softmax Lipschitz, $\|\pi' - \pi\|_\infty \le \|h\|_2 \epsilon$; hence $|\Delta H_1| \le L_H \cdot 2\|h\|_2 \epsilon = O(\epsilon)$.

Step 2: Gradient flow. $\partial \mathcal{L}/\partial x_l = \partial \mathcal{L}/\partial x_L \cdot \prod_{j=l}^{L-1} (I + \partial(\Delta A_j + \Delta F_j)/\partial x_j)$. The identity path bounds $\|\partial \mathcal{L}/\partial x_l\| \le G_l$ (A4).

Step 3: Activation perturbation. $\|\Delta H_l\|_2 \le \eta G_l J_l$ where J_l is the Jacobian norm.

Step 4: Singular value bounds (Weyl). $|\sigma_i' - \sigma_i| \le \|\Delta H_l\|_2 \le \eta G_l J_l \equiv \epsilon_l$.

Step 5: Effective rank bound. $|\Delta \mathrm{erank}| \le L_{\mathrm{erank}} \sum_i |\hat{\sigma}_i' - \hat{\sigma}_i| \le 2 L_{\mathrm{erank}} \eta G_l J_l / C$.

Step 6: KL attenuation. Under (A2), $G_l \le G_{\max} \sqrt{\delta(\beta, \epsilon)}$, so $f \le 2 L_{\mathrm{erank}} \eta G_{\max} \sqrt{\delta(\beta, \epsilon)} J_l / C$, with $f \to 0$ as $\beta \to \infty$.

Step 7: Gap. $|\Delta \mathrm{erank}| / |\Delta H_1| \le L_{\mathrm{erank}} G_l J_l / (C \|h\|_2) \ll 1$ when η small, $\beta > 0$, C large. ■

B Proof of Theorem 2 (CKA Blind Spot)

Setup. Let $X \in \mathbb{R}^{n\times d}$ be centered and $X' = X + \Delta X$. Linear CKA is $\mathrm{CKA}(X, X') = \|X^\top X'\|_F^2 / \big(\|X^\top X\|_F \, \|X'^\top X'\|_F\big)$. Write $G = X^\top X$ and $P = X^\top \Delta X$, both in $\mathbb{R}^{d\times d}$.

Step 1: Numerator. $\|X^\top X'\|_F^2 = \|G + P\|_F^2 = a + 2b + \|P\|_F^2$, where $a = \|G\|_F^2$ and $b = \langle G, P\rangle_F$.

Step 2: Denominator. $X'^\top X' = G+P+P^\top+\Delta X^\top \Delta X = G+S+\Delta X^\top \Delta X$ with $S = P+P^\top$ symmetric. Since G is symmetric, $\langle G, P^\top\rangle_F = \langle G, P\rangle_F$, so $\langle G, S\rangle_F = 2b$. Thus $\|X'^\top X'\|_F^2 = a + 4b + O(\epsilon^2)$, hence $\|X'^\top X'\|_F = \sqrt{a}\,\big(1 + 2b/a + O(\epsilon^2)\big)$.

Step 3: Exact cancellation of the linear term. The full denominator is $\|X^\top X\|_F \, \|X'^\top X'\|_F = \sqrt{a}\cdot\sqrt{a}\,(1 + 2b/a + O(\epsilon^2)) = a + 2b + O(\epsilon^2)$. The numerator is $a + 2b + \|P\|_F^2$. The linear terms $2b$ are *identical*, so they cancel exactly:

$$\mathrm{CKA} = \frac{a + 2b + \|P\|_F^2}{a + 2b + O(\epsilon^2)} = 1 + \frac{\|P\|_F^2 - O(\epsilon^2)}{a} = 1 - \Theta(\epsilon^2).$$

No first-order term survives. The cancellation is exact—not a smallness assumption—because CKA is a ratio in $[0, 1]$ whose numerator and denominator share the linear growth.

Step 4: Order and constant. $\|P\|_F = \|X^\top \Delta X\|_F \le \|X\|_2 \|\Delta X\|_F = O(\epsilon)$, giving $1 - \mathrm{CKA} = \Theta(\epsilon^2)$. The leading constant $\lim_{\epsilon\to 0}(1 - \mathrm{CKA})/\epsilon^2$ depends on G and the perturbation direction; a deterministic check yields ≈ 0.67 (not 2) for isotropic noise (Appendix I). The claim of a universal constant 2 in earlier drafts is therefore withdrawn; the order-of-magnitude statement is unchanged.

Step 5: Linear probe sensitivity. For $\Delta X = \delta\, uv^\top$ with unit u, v, $\|\Delta X\|_F = \delta$ arbitrarily small. The classifier $w = v$ gives $w^\top x_i' = w^\top x_i + \delta u_i$; if u_i has the sign of the class, it separates perfectly with margin $\delta \cdot \min_i |u_i|$. Probe sensitivity is $O(\epsilon)$ (linear); CKA sensitivity is $O(\epsilon^2)$ (quadratic), a $\sim 1/\epsilon$ gap.

Step 6: Weight-space connection. $\Delta X \approx X \Delta W^\top$ (first-order Taylor), so $\|\Delta X\|_F / \|X\|_F \le \|\Delta W\|_F$. With $\|\Delta W\|_F / \|W\|_F \approx 3.3 \times 10^{-4}$ (Phase 49), $1 - \mathrm{CKA} = \Theta(\epsilon^2) \sim 10^{-7}$, consistent with the empirical reading at float32 precision (the deviation is at the representable resolution). ■

C Protocol Audit Details

The evaluation protocol evolved across phases: Phases 1–22 used no `\boxed{}` instruction with lenient parsing; Phase 23+ added `\boxed{}` and the chat template; Phase 26 froze the faithful v2 protocol. The interaction between the parsing function and training-induced format changes creates the phantom narrowing signal. The protocol-effect/training-effect comparison is quantified in Section 2.

D Fixed-Dictionary SAE Confound Details

Six independently trained SAEs (one per condition) all show $\Delta\text{alive} \le 8$ between baseline and trained models on the same distribution, vs. $\Delta\text{alive} = +3278$ under the fixed-dictionary protocol. The ratio $R_{\mathrm{RLVR/DS}} = |\Delta S_{\text{RLVR}}|/|\Delta S_{\text{drift}}| = 3278/11317 \approx 0.29 < 1$ confirms drift-dominated statistics: any reported "feature recruitment" under a fixed dictionary is confounded.

E Estimator Noise Details

The unseeded 5000-row subsampling estimator has SD ≈ 8.7 (measured over 100 seeded runs). Two independent unseeded runs on the same data produced $\Delta\mathrm{erank} = -11.64$ and $+1.41$—a sign reversal. The deterministic full-pool SVD gives $\Delta\mathrm{erank} = +0.02$, confirming geometric invariance. Recommendation: deterministic full-pool SVD, or 100-seed SD when subsampling is unavoidable.

F C1–C5 Validity Checklist

	Condition	Detection
C1	Protocol consistency	Protocol audit
C2	Dictionary calibration	$R_{\mathrm{RLVR/DS}} < 1$
C3	Adaptive validation	Per-condition SAE retraining
C4	Estimator determinism	Full-pool SVD or 100-seed SD
C5	Multi-seed robustness	≥ 3 training seeds

G Secondary Experiments and Related Work

Temporal dynamics (Phase 36). CKA > 0.9997 at every checkpoint $\{100, 500, 1000, 1500, 2000\}$ and layer; the Gap is not an endpoint artifact. (Defensive.)

Statistical power (Phase 38). Bootstrap null gives MDES $= 0.0017$ at 80% power ($n = 2{,}000$ tokens); sensitivity calibration shows CKA drops below 0.99 only at perturbations ≥ 0.2 (rotation) or ≥ 0.5 (noise/rank), far above any observed RLVR effect (< 0.02). (Defensive.)

Explicit erank/entropy regularization (Phase 37). Across 5 variants $\times$ 3 layers, mean CKA $= 0.999998$ and zero significant neurons; max $|\Delta\mathrm{erank}| = 1.34$ (layer 24, $\lambda = 0.1$), only 0.5% of baseline—consistent with Theorem 1 (L3 intervention moves geometry, L1/L2 do not).

Gradient flow (Phase 39). 75.5% of gradient norm concentrates in layers 0–8, yet CKA > 0.999: the gap is not gradient absence but gradient direction living in the representation null space.

Multi-layer neurons (Phase 34) and scale (Phase 32). The $50.4\times$ neuron discrimination holds across all layers (mean $469.9\times$ KS ratio), and the CKA pattern is qualitatively identical at 0.5B (higher overall CKA, consistent with capacity dependence).

SAE feature stability (Phase 40). 0% of $12{,}288$ features differ at layers 3/14 for GRPO (DPO: $0\%/4.3\%$); layer 24 shows 31.8% (GRPO) / 94.1% (DPO), a progressive pattern CKA still reports as identity (≥ 0.999996).

Weight space (Phase 41) and multi-task (Phase 42). Minimum weight cosine 0.99990 across 7 variants; MATH representations give CKA $= 0.999998$, and a cross-task positive control (GSM8K vs. MATH prompts) drops CKA to 0.0004—confirming the null is training-specific, not a general insensitivity.

Related work. CKA Kornblith et al. (2019) is the standard representation metric, complementing SVCCA and orthogonal Procrustes; RankMe Garrido et al. (2023) estimates effective rank and is susceptible to Pitfall 3; Li et al. Li et al. (2025) use RankMe on RLVR (potentially affected); Minder et al. Minder et al. (2025) use crosscoders for chat-tuning diffing (potentially affected by the fixed-dictionary confound); Cui et al. Cui et al. (2025) and Wu et al. Wu et al. (2025) assume RLVR measurably changes representations, which our Gap theorem qualifies; the "Sober Look" work Various (2025) raised reproducibility concerns that we extend to interpretability methods.

H Reproducibility

All scripts, cached activations, and checkpoints are available; the audit toolkit implements C1–C5. All spectral statistics use deterministic full-pool SVD; bootstrap CIs use $n_{\text{boot}} = 10{,}000$ with fixed seed; neuron KS tests use Bonferroni correction; the paradox ratio uses fp32 arithmetic to avoid numerical artifacts; head-swap experiments use untied LM-head weights.

I NUMERICAL VERIFICATION OF THEOREM 2

Theorem 2 makes two claims we verified deterministically (seed 42, $X \in \mathbb{R}^{200\times100}$ standard normal; script `theorem_verify_cka_blindspot.py`): (i) $1 - \text{CKA}$ is purely quadratic in ϵ, and (ii) the leading constant is not universal. A log–log fit of $1 - \text{CKA}$ versus $\epsilon \in \{10^{-3}, \ldots, 10^{-1}\}$ gives slope 1.991 (quadratic, not linear). The effective constant $(1 - \text{CKA})/\epsilon^2$ converges to ≈ 0.67 for isotropic noise—not 2—and is essentially identical whether the perturbation ΔX lies inside or orthogonal to the column space of X (both ≈ 0.67), confirming the first-order term cancels exactly rather than by smallness. The earlier draft's claim $\text{CKA} \approx 1 - 2\epsilon^2$ and "matches to six decimals" is therefore withdrawn in favor of the order-of-magnitude statement $\text{CKA} = 1 - \Theta(\epsilon^2)$.