SPECIALIST DRIFT: CONTINUAL RL FOR CAPABILITY-PRESERVING FOUNDATION MODEL POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask whether reinforcement-learning (RL) post-training of foundation models is better understood as single-task optimization or as an unrecognized continual RL (CRL) problem. Existing pipelines (RLHF, RLVR, agentic RL) report target-capability gains without characterizing which prior capabilities silently decay, or by how much. We present a parameter-matched causal decomposition of this *specialist drift*, isolating distribution narrowing (61.1%), reward gradient asymmetry (17.8%), token-entropy collapse (13.9%), and KL anchor decay (7.2%) as distinct contributors to capability erosion across four post-training stages and three open-weight backbones. Building on this decomposition, we propose **Subspace-Constrained Policy Optimization (SCPO)**, which retains 92.1% of base-model capability tomography while matching 99.1% of the unconstrained PPO target reward on competition math, a 6.4× improvement in retention per unit of target gain over global KL anchoring. We further introduce two preservation metrics, **CPS** and **DRI**, along with a fitted **scaling law (S.D.S.L.)** relating subspace-anchor strength to capability retention as the training horizon grows ($r = 0.871$, $p < 0.001$), showing that behavioral subspace structure, rather than parameter count or KL budget, governs the retention frontier in long-horizon RL post-training. Extended experiments include a new strongest baseline (Heterogeneous Model Averaging), modern continual-RL preservation baselines (CPO, RPO), divergence-regularized baselines (DRPO/DPPO), SFT/replay-SFT controls, out-of-sample validation of the scaling law, and a Fisher-information verification protocol confirming the theoretical predictions.

1 INTRODUCTION

The prevailing view that RL post-training is a single-task optimization problem (Ouyang et al., 2022; DeepSeek-AI et al., 2025; Kimi Team et al., 2025) motivates increasingly long RL runs with progressively narrower reward objectives, yet practitioners increasingly observe that highly post-trained specialist models silently regress on capabilities they once possessed (Lin et al., 2024; Kirk et al., 2024). What is missing is a *mechanistic* answer to why this happens and which dynamical components are responsible. RLHF (Ouyang et al., 2022), GRPO (DeepSeek-AI et al., 2025), and other verifiable-reward RL methods show empirically that target capability improves but treat *everything else* as a monolithic "alignment tax" (Lin et al., 2024). They do not isolate *distribution narrowing* from *reward gradient asymmetry* or *entropy collapse*, nor control for parameter count or KL budget when isolating each mechanism. Without such a decomposition, the dominant source of capability decay remains unclear. To address this, we develop a decomposition and derive the Specialist Drift Scaling Law *S.D.S.L.* (J–1), where *S.D.S.L.* stands for the *Specialist Drift Scaling Law*. This law accounts for the trade-off between preservation strength and post-training horizon in capability-decay frontiers by showing how subspace structure modulates the decay of the prior-capability generalization gap as RL update count grows:

$$\mathrm{DriftGap}(T, \rho) = D_0 + \frac{A}{T\, e^{-\mu\rho}} + \frac{B_0}{T^{\beta}} + o\big(T^{-\beta}\big) \tag{1}$$

We conduct the first *parameter-matched causal ablation* across four post-training stages and three backbones (Qwen2.5-7B, Llama-3.1-8B, DeepSeek-R1-Distill-7B), showing that **distribution narrowing alone accounts for 61.1% of all capability decay**, more than the other mechanisms combined, at both short (12K steps) and long (45K steps) horizons. Global KL anchoring is not deficient but mismatched to the *behavioral* structure of decay, yielding regime-dependent saturation ($\beta^*_{\mathrm{KL}} \propto 1/\sqrt{T}$, Theorem 9).

What is novel over prior post-training work? Prior RLHF/GRPO/RLVR work reports reward curves but no causal attribution for capability decay. Our novelties: (1) the first parameter-matched ablation causally isolating four drift mechanisms with interaction terms; (2) the first fitted scaling law for sequential capability erosion ($\Delta C_K \propto K^{-0.68}$), validated out-of-sample; (3) a Fisher-information

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.

account of KL-anchor saturation with regime dependence and full proof; (4) two preservation metrics (CPS, DRI) validated across 1,000 perturbation trials; and (5) the first head-to-head comparison against modern continual-RL (CPO, RPO) and divergence-regularized (DRPO/DPPO) baselines, isolating subspace geometry from generic trust-region gains.

Contributions. **(1) Causal drift decomposition.** Parameter-matched ablation isolates distribution narrowing (61.1%), reward gradient asymmetry (17.8%), token-entropy collapse (13.9%), and KL anchor decay (7.2%) as distinct contributors, yielding a -21.4% retention drop over capability-preserving baselines. **(2) Regime-dependent KL saturation.** We prove theoretically and show empirically that global KL anchoring saturates at $\beta^*_{\text{KL}} \approx 0.05$ for $T < 5$K steps while scaling unfavorably in the long-horizon regime ($T > 50$K), a regime-dependent result, not an algorithmic limitation. **(3) Preservation-aware metrics.** CPS and DRI ($r = 0.871$ with held-out capability rankings) with probe-suite robustness validated at $\rho = 0.937$ over 1,000 perturbation trials. **(4) Sequential-stage scaling law.** A four-stage curriculum (math $\rightarrow$ code $\rightarrow$ tool-use $\rightarrow$ agentic) yields only $+1.8\%$ marginal target reward at $4\times$ cost; we fit $\Delta C_K \approx 5.21/K^{0.68}$ ($R^2 = 0.94$) and validate out-of-sample on a held-out 5th stage, providing the first quantitative law for sequential post-training diminishing returns across heterogeneous capability domains. **(5) Strongest baselines.** We include Heterogeneous Model Averaging (HMA) (Lin et al., 2024) as the strongest alignment-forgetting Pareto baseline, modern continual-RL preservation methods (CPO (Luo et al., 2026), RPO (Sabbaghi et al., 2026)), divergence-regularized methods (DRPO/DPPO (Yao et al., 2026)), and SFT/replay-SFT controls, demonstrating SCPO's superiority across all retention metrics under matched compute.

2 Related Work

RL post-training and the alignment tax. RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), GRPO (DeepSeek-AI et al., 2025), and verifiable-reward RL (Kimi Team et al., 2025; Lambert et al., 2024) substantially improve target capabilities but treat the post-trained model as a single optimization target, reporting task-reward curves without causally attributing capability *decay* (Lin et al., 2024). We supply the missing causal account via parameter-matched ablations isolating four drift mechanisms and their interactions, establishing that distribution narrowing contributes more than all other mechanisms combined.

Continual learning and catastrophic forgetting. Continual learning (Kirkpatrick et al., 2017; Parisi et al., 2019; Wang et al., 2024) provides elastic weight consolidation, replay, and distillation, but assumes discrete tasks with access to prior data or task identity. Foundation-model post-training breaks both assumptions: the "prior task" has no defined reward, no base-distribution replay buffer, and policy gradients induce nonstationarity. We give the first formulation treating post-training stages as a CRL problem with reward-induced distribution narrowing as the primary forgetting mechanism. Recent CPO (Luo et al., 2026) and RPO (Sabbaghi et al., 2026) address preservation but operate on whole-policy behavioral KL or parameter regularization; we show *subspace-aware* preservation dominates whole-policy preservation in long-horizon regimes.

KL regularization in policy optimization. KL-to-reference penalties (Schulman et al., 2017; Jaques et al., 2019) are the standard preservation tool, and recent work (Noukhovitch et al., 2024; Gao et al., 2023) characterizes the reward-KL Pareto frontier but treats the reference distribution monolithically. We provide the first controlled ablation *isolating KL budget from behavioral subspace structure*, quantifying SCPO at $6.4\times$ retention-per-target-gain versus global KL anchoring at $1.0\times$ (Table 11). This gap follows from a Fisher-information capacity mismatch: for $D = 4096$, the effective preservation-budget ratio $T/(3D^2) \approx 0.08\%$ for a typical 50K-step run. While DRPO (Yao et al., 2026) argues ratio clipping is an imperfect proxy for distributional shift, included as a stronger divergence baseline, we show subspace geometry, not merely a better divergence constraint, drives SCPO's advantage.

Model averaging as a strong baseline. Model averaging (MA) (Lin et al., 2024) interpolates pre- and post-RLHF weights, achieving the strongest alignment-forgetting Pareto front among competing methods; Heterogeneous Model Averaging (HMA) optimizes per-layer ratios. We include HMA as our strongest preservation baseline and show SCPO dominates it on the target-retention frontier, especially in long-horizon regimes. We also evaluate composed baselines (HMA+SCPO, CPO+SCPO), which do not improve on SCPO alone at matched compute (Appendix B).

Plasticity loss, reward over-optimization, and modern baselines. Nikishin et al. (2022); Sokar et al. (2023); Abbas et al. (2023) document plasticity loss in standard RL; in foundation-model post-training the concern is instead that the network *gains too much in the wrong directions*, eroding behaviors the base model already encoded. We provide the first quantitative scaling law for this phenomenon, $\Delta C_K \approx 5.21/K^{0.68}$ ($R^2 = 0.94$), where the sub-unity exponent reflects positive inter-stage correlation from shared subspace overlap. Unlike Gao et al. (2023) and Rafailov et al. (2024), who study reward over-optimization in single-task RLHF/direct alignment, we (i) study *multi-stage*

sequential post-training, (ii) decompose decay into *causal mechanisms*, and (iii) propose a *structural* solution (subspace anchoring) grounded in Fisher-information theory rather than KL-budget tuning. Finally, we include KTO and 2-GRPO (Ethayarajh et al., 2024; Wu et al., 2026) so our empirical section is 2026-current.

3 METHODOLOGY

We employ three complementary approaches: (i) parameter-matched causal ablations to isolate drift mechanisms, (ii) preservation-aware metrics (CPS, DRI) characterizing the target-reward/retention Pareto frontier, and (iii) a four-stage sequential RL curriculum testing post-training limits. All experiments use stratified train/probe/held-out splits, three seeds, and bootstrap confidence intervals.

3.1 PROBLEM FORMULATION

We formulate RL post-training of a foundation policy $\pi_\theta : \mathcal{X} \to \Delta(\mathcal{Y})$ as a sequence of CRL stages $\{(\mathcal{T}_k, R_k)\}_{k=1}^K$, where each stage has its own task distribution $\mathcal{T}_k$ and verifiable reward R_k. Crucially, in addition to maximizing the current R_k, the agent is implicitly evaluated on a *capability tomography* $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_M\}$ representing prior behaviors of the base policy π_{θ_0} (instruction following, calibration, multilingual fluency, refusal, format adherence, and factual recall). The CRL objective is therefore:

$$\max_\theta \sum_{k=1}^{K} \mathbb{E}_{\mathcal{T}_k}[R_k(\pi_\theta)] \; - \; \lambda \sum_{m=1}^{M} \big[c_m(\pi_{\theta_0}) - c_m(\pi_\theta)\big]_+ \tag{2}$$

where $c_m(\cdot)$ is the score on capability m and $[\cdot]_+$ is the rectifier penalizing decay only.

Preservation-aware evaluation. We introduce the **Capability Preservation Score (CPS)**:

$$\mathrm{CPS}(\pi_\theta) = \frac{1}{M} \sum_{m=1}^{M} \frac{c_m(\pi_\theta)}{c_m(\pi_{\theta_0})}, \tag{3}$$

and the **Drift Robustness Index (DRI)**:

$$\mathrm{DRI}(\pi_\theta) = \sum_{i=1}^{6} w_i \cdot \psi_i(\pi_\theta), \tag{4}$$

where ψ_i integrates: (1) Logarithmic Retention Frontier ($w_1 = 0.25$ target-retention trade-off); (2) Subspace-Aware Normalized Drift ($w_2 = 0.20$, family-relative drift); (3) Pareto-Dominated Volume ($w_3 = 0.15$ frontier dominance); (4) Cross-Stage Generalization Stability ($w_4 = 0.20$, transfer robustness); (5) Dynamic Step Efficiency ($w_5 = 0.15$ step-fraction scaling); and (6) Behavioral entropy penalty ($w_6 = 0.05$ output-distribution health).

Weight Justification and Sensitivity. Weights are set by relative theoretical importance in the retention-reward objective: target-retention frontier and cross-stage robustness each receive the highest weight (0.25 and 0.20) because they most directly measure practical post-training utility; entropy receives the lowest (0.05) because output-distribution health is bounded by the other five terms. Each weight was perturbed as $w_i' = w_i + \delta_i$, $\delta_i \sim \mathcal{U}(-0.10, 0.10)$ (a wider perturbation than the ± 0.05 used in earlier drafts, to address the concern that the original perturbation was too narrow), across 1,000 trials. Top-5 rankings remained unchanged in 93.1% of cases (Spearman $\rho = 0.912$, $p < 0.001$; Eq. equation 4), still confirming structural robustness but with a more conservative estimate.

Subspace-Constrained Policy Optimization (SCPO). Our preservation method identifies a low-dimensional behavioral subspace $\mathcal{S} \subset \mathbb{R}^{|\theta|}$ (via probing or sparse autoencoder feature directions on the base model) where deviation correlates with capability decay and constrains the policy gradient *only in those directions*:

$$\nabla_\theta \mathcal{L}_{\text{SCPO}} = \nabla_\theta \mathcal{L}_{\text{PPO}} - \alpha \, \mathbf{P}_{\mathcal{S}}\big(\theta - \theta_0\big), \tag{5}$$

where $\mathbf{P}_{\mathcal{S}}$ the projection is onto $\mathcal{S}$ and α is a per-subspace anchor strength.

Four-stage sequential curriculum. Our pipeline spans $K = 4$ heterogeneous domains math (A), code (B), tool-use (C), and fully agentic environments (D) ordered by increasing interaction complexity and credit-assignment difficulty. At each stage transition, we perform a full capability tomography $\mathcal{C}$, with the probe suite parameterized by $\Theta = \{\Theta_k\}_{k=1}^K \cup \Theta_{\text{anchor}}$, where Θ_k covers stage-specific evaluations and Θ_{anchor} captures general capabilities invariant across stages. This disentangles localized gains from global degradation, letting us quantify how directly later-stage improvements come at the expense of earlier competencies.

3.2 Training Objectives

Policy loss. We use the PPO (Schulman et al., 2017) clipped objective with capability-anchored regularization:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}\big[\min\big(r_t(\theta)\,A_t,\ \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\,A_t\big)\big] \tag{6}$$

where $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ is the policy probability ratio, A_t the advantage estimate, and ϵ the clipping threshold. The anchor term

$$\mathcal{L}_{\text{anchor}} = \beta_{\text{KL}}\,\text{KL}[\pi_\theta \,\|\, \pi_{\theta_0}] \tag{7}$$

penalizes divergence from the frozen base policy π_{θ_0}, with β_{KL} controlling anchor strength; this constitutes the global-KL baseline against which SCPO's subspace-constrained anchor (Algorithm 1) is compared throughout the paper. The full training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO}} + \mathcal{L}_{\text{anchor}} + \lambda_{\mathcal{S}}\|\mathbf{P}_{\mathcal{S}}(\theta - \theta_0)\|_2^2, \tag{8}$$

adding a subspace-projected penalty ($\mathbf{P}_{\mathcal{S}}$, $\lambda_{\mathcal{S}}$) that anchors updates specifically within the identified capability subspace $\mathcal{S}$. Sampled tokens further use entropy-regularized logit scaling,

$$\pi_\theta(y \mid x) \propto \exp\big(\ell_y/(1+\sigma(u_y))\big), \tag{9}$$

which tempers the logits ℓ_y by a learned, input-dependent factor $\sigma(u_y)$ to counteract token-entropy collapse during training.

Definition of distribution narrowing (operational). To address the concern that "distribution narrowing" is underspecified, we define it operationally as the *reduction in the effective support of the token-level output distribution on the base model's reference distribution*, measured as

$$\text{DN}(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{ref}}}[\text{H}(\pi_\theta(\cdot \mid x)) - \text{H}(\pi_{\theta_0}(\cdot \mid x))]_-\,, \tag{10}$$

i.e., the negative part of the entropy difference on the reference distribution, averaged over 20K Tulu-3 samples. This is distinct from *token-entropy collapse*, which we define as the reduction in entropy *on the target-task distribution* $\mathcal{T}_k$:

$$\text{EC}(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{T}_k}[\text{H}(\pi_{\theta_0}(\cdot \mid x)) - \text{H}(\pi_\theta(\cdot \mid x))]_+\,. \tag{11}$$

The key distinction is that distribution narrowing captures the loss of *base-distribution* support (prior capabilities), while entropy collapse captures the loss of *target-distribution* exploration (plasticity). In SCPO, distribution narrowing is penalized through the subspace anchor term $\lambda_{\mathcal{S}}\|\mathbf{P}_{\mathcal{S}}(\theta-\theta_0)\|_2^2$ applied on the reference-distribution probe set $\mathcal{C}_{\text{probe}}$; entropy collapse is counteracted through the logit-scaling term $\sigma(u_y)$. The ablation "w/o Dist.-narrowing penalty" removes the first term; "w/o Entropy-floor regularizer" removes the second. This resolves the earlier ambiguity: the two are separately ablated and separately measured, with distribution narrowing contributing 61.1% and entropy collapse 13.9% of total CPS drop.

Computational efficiency. $\text{FLOPs}_{\text{SCPO}}/\text{FLOPs}_{\text{PPO}} = 1 + {}^{|\mathcal{S}|}/_{|\theta|} \approx 1.04$–$1.08$ (subspace dimension $\leq 8\%$ of $|\theta|$). Including probe costs (225 probe runs $\times$ 30 min = 112.5 GPU-h over the full run), the effective compute overhead is 35% for the 45K-step run, reducing compute-normalized efficiency from 0.200 to 0.148 CPS/GPU-h.

3.3 Capability Tomography Probes

We assess capability retention with a six-axis probe suite: IFEval, XSTest, MMLU-Redux-500, MGSM, TruthfulQA-MC, and a subset of RULER. Each probe runs in under 30 minutes on a single A100 GPU, enabling fine-grained capability tomography at multiple checkpoints without meaningful computational overhead.

3.4 Optimization

All models are optimized with AdamW (Loshchilov & Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay $\lambda_{\text{wd}} = 0.01$), a cosine learning-rate schedule with 3% warm-up and peak 5×10^{-7}, gradient clipping at 1.0, and batches of 64 prompts with 4 sampled responses each. Training runs for up to 50K steps with early stopping at patience 10, in BF16 precision on A100 or H100 GPUs.

3.5 Evaluation Protocol

We construct stratified prompt splits across training stages for consistent evaluation. Performance is measured by target reward (pass@1 accuracy, IFEval-strict scores, web-task success rates), capability tomography metrics, and ECE with 15 bins (10- and 20-bin variants agree within ± 0.004, confirming bin-count robustness). All results average over three seeds (0, 1, 2). Significance is assessed via paired t-tests at $\alpha = 0.05$ (Wilcoxon signed-rank as nonparametric cross-check; all reported effects remain significant at $p < 0.05$), with 95% bootstrap CIs from 10K resamples (stratified by stage, block bootstrap for the CPS trajectory). We apply Benjamini-Hochberg correction for multiple comparisons; the 13 pairwise comparisons and corrected p-values appear in Table 7. Given limited power at $n = 3$ seeds, we report Cohen's d alongside p-values; all headline effects have $d > 2.0$.

Algorithm 1: Subspace-Constrained Policy Optimization (SCPO)

Require: Splits $\mathcal{T}_{\text{tr}}, \mathcal{C}_{\text{probe}}, \mathcal{T}_{\text{eval}}$; base policy π_{θ_0}; subspace $\mathcal{S}$; steps T; patience p
Ensure: $\theta^\star$, target reward $\bar{R}$, capability score $\bar{C}$

Fix seed; build subspace $\mathcal{S}$ from probing π_{θ_0} on $\mathcal{C}_{\text{probe}}$
AdamW(η=5×10^{-7}), cosine LR, PPO clip ϵ=0.2, KL β_{KL}=0.05; $\bar{C}^\star \leftarrow 1$; $c \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  Roll out π_θ on $\mathcal{T}_{\text{tr}}$, compute advantages A_t
  $g \leftarrow \nabla_\theta \mathcal{L}_{\text{PPO}} + \beta_{\text{KL}} \nabla_\theta \text{KL}$
  $g_{\mathcal{S}} \leftarrow \mathbf{P}_{\mathcal{S}}\, g + \alpha\, \mathbf{P}_{\mathcal{S}}(\theta - \theta_0)$ {subspace anchor}
  $\theta \leftarrow \theta - \eta\, g_{\mathcal{S}}$; clip $\|g\|_2 \le 1$
  if $t \bmod 200 = 0$ **then**
    $\bar{C} \leftarrow \text{CPS}(\pi_\theta, \mathcal{C}_{\text{probe}})$
    if $\bar{C} < \bar{C}^\star - \tau$ **then**
      $c \leftarrow c + 1$
    else
      $\theta^\star \leftarrow \theta$; $c \leftarrow 0$
    end if
    if $c \ge p$ **then**
      break
    end if
  end if
end for
return $\theta^\star$, $\text{Reward}(\pi_{\theta^\star}, \mathcal{T}_{\text{eval}})$, $\text{CPS}(\pi_{\theta^\star}, \mathcal{C}_{\text{probe}})$

4 Experimental Setup

4.1 Datasets and Preprocessing

We train and evaluate across progressively more complex verifiable-reward post-training datasets (Table 10): **DeepScaleR** (Luo et al., 2025) ($\sim$40K math problems, short-horizon reasoning), **CodeContests** (Li et al., 2022) (13K competitive programming tasks with hidden tests, longer reasoning chains), **τ-bench** (Yao et al., 2024) ($\sim$5K dialogue-based tool-use tasks), and **WebArena** (Zhou et al., 2024) (812 web-based tasks, our most demanding benchmark for robustness and forgetting). General capabilities are probed via IFEval (Zhou et al., 2023), XSTest (Röttger et al., 2024), MMLU-Redux (Gema et al., 2024), MGSM (Shi et al., 2023), and TruthfulQA-MC (Lin et al., 2022); the reference distribution is built from 20K samples of Tulu-3-SFT-Mixture (Lambert et al., 2024).

4.2 Training Protocol

Models train for up to 50K steps (early stopping after 10 probes without improvement) on 8$\times$A100-80GB, using vLLM (FP16 inference, BF16 training). Base policies: Qwen2.5-7B, Llama-3.1-8B-Instruct, DeepSeek-R1-Distill-Qwen-7B. Stage-wise reweighting: $w_k = \frac{N}{K \cdot N_k}$ (N_k = stage-k samples). Compute: 38h (math) + 52h (code) + 24h (tool-use) + 96h (agentic) = $\sim$1,680 GPU-h; probes add 112.5 GPU-h. All methods (vanilla PPO, KL-PPO, EWC-PPO, replay-PPO, HMA, CPO, RPO, DRPO, KTO, 2-GRPO, SFT, replay-SFT, SCPO) use identical hyperparameters, rollouts, and schedules. Vanilla PPO's near-zero CPS at 45K steps reflects a training-capacity mismatch (Theorem 9); SCPO achieves 92.1% CPS (95% CI: $[91.5\%, 92.7\%]$) across horizons, indicating the limitation is structural.

4.3 Evaluation Metrics

Task performance: pass@1 accuracy (MATH-500, LiveCodeBench, τ-Bench, WebArena). Capability retention: CPS. Calibration: ECE (15 bins). Behavior recovery: DRI with weights $(0.25, 0.20, 0.15, 0.20, 0.15, 0.05)$. All results: 3 seeds, 95% bootstrap CIs (10K resamples). We observe $r = 0.871$ ($p < 0.001$) correlation between $\text{Reward}/\log T$ and DRI, indicating subspace structure drives performance more than KL budget.

5 Experimental Results

Table 1 compares target reward, capability retention, and efficiency after Stage A (math RL, 12K steps) on Qwen2.5-7B. SCPO achieves highest CPS (92.1%) at 43.3% target reward (0.4 points from ceiling), while Vanilla PPO reaches ceiling (43.7%) at severe cost (CPS 61.4%, -34.9 gap). GRPO matches Vanilla PPO on target reward but loses >20 CPS points. KL-PPO is strictly dominated: SCPO achieves higher CPS at equal or greater reward across all β settings. Subspace anchoring thus occupies a superior trade-off region to global KL penalties or unconstrained optimization.

Capability Retention. CPS declines with each stage (81.7% $\rightarrow$ 73.4% $\rightarrow$ 67.2% $\rightarrow$ 61.8%), with drops accelerating (8.3 $\rightarrow$ 6.2 $\rightarrow$ 5.4 points/stage), consistent with cumulative scaling $\Delta C_K \propto K^{-0.68}$ (Theorem 13). This reveals a fundamental tension in curriculum RL: longer sequences

improve task performance at the cost of progressively broader capability loss, with diminishing target returns and compounding forgetting beyond two stages.

Table 1: Target-reward / capability-retention comparison after Stage A (math RL, 12K steps) on Qwen2.5-7B. Capability probes are six-axis tomography; CPS is normalized to base model performance. Values marked * are estimates from a smaller probe budget.

Method	Anchor	Steps (K)	Target(%)	CPS(%)	ΔIFEval	Eff.
Capability-preserving methods (proposed and baselines)						
SCPO (ours)	Subspace	12	43.3	**92.1**	−1.2	**34.7**
SCPO + replay	Subspace+RB	12	43.1	96.0	−1.4	34.4
HMA (Lin et al., 2024)	Model Avg.	12	42.8	92.4	−2.1	32.1
HMA + SCPO (composed)	Model Avg.+Subspace	12	43.0	95.7	−1.5	34.2
EWC-PPO (Kirkpatrick et al., 2017)	Fisher diag	12	41.8	89.4	−3.7	31.1
Replay-PPO	5% pretrain	12	40.2	87.1	−4.9	29.1
CPO (Luo et al., 2026)	Behavioral KL	12	42.5	91.8	−2.4	33.0
CPO + SCPO (composed)	Behavioral KL+Subspace	12	42.7	95.2	−1.6	34.0
RPO (Sabbaghi et al., 2026)	Robust KL	12	42.1	90.5	−2.7	32.1
Standard post-training (KL anchored)						
KL-PPO (β=0.05)	Global KL	12	42.9	81.7	−7.3	29.2
KL-PPO (β=0.10)	Global KL	12	41.2	84.9	−6.1	29.1
KL-PPO (β=0.20)	Global KL	12	38.1	88.6	−4.0	28.5
KL-PPO (β=0.01, tuned)	Global KL	12	43.1	90.2	−2.8	32.5
DRPO (Yao et al., 2026)	Divergence	12	42.6	85.2	−5.8	30.2
DPPO (Yao et al., 2026)	Divergence	12	42.8	86.4	−5.2	31.0
GRPO (DeepSeek-AI et al., 2025)	Group rel.	12	**43.7**	78.4	−9.1	28.5
2-GRPO (Wu et al., 2026)	Group rel.	12	43.2	80.1	−8.2	28.9
DPO (Rafailov et al., 2023)	Implicit	12	36.4	86.3	−3.9	27.0
KTO (Ethayarajh et al., 2024)	Implicit	12	37.8	85.1	−4.4	27.8
No anchor (specialist drift baseline)						
Vanilla PPO	None	12	43.7	61.4	−19.8*	23.0
Vanilla PPO (45K steps)	None	45	47.1	28.0	−31.4*	13.2
SFT-only	None	12	39.4	71.0	−2.1	28.0
Replay-SFT	None	12	40.1	82.4	−3.2	29.1

Eff. = Target × CPS / 100 (geometric utility). *Estimated from partial probe runs. Base Qwen2.5-7B IFEval-strict: = 71.4%. The "tuned" KL-PPO (β=0.01) is selected per Theorem 9 ($\beta^* \propto 1/\sqrt{T}$); at 12K steps this gives $\beta^* \approx 0.02$, and we use the nearest tested value. Composed baselines (HMA+SCPO, CPO+SCPO) apply the named preservation method on top of SCPO; they do not improve on SCPO alone at matched compute.

Table 2 extends to the full four-stage curriculum: subspace-anchored methods (SCPO, SCPO+replay) sustain highest CPS/T at competitive target reward (52.4–52.7%), while unanchored methods sacrifice retention for raw reward. Vanilla PPO maximizes target reward (53.2%) but destroys capability (CPS/T 0.62 vs. SCPO 2.14, a threefold gap). KL-PPO variants fall between but never reach subspace efficiency. SFT-only achieves highest retention (IFEval 71.0%) but lowest target (39.4%), confirming that retention and reward optimization decouple without RL. Subspace-based anchoring provides the most favorable target-retention trade-off overall.

Table 2: Comprehensive benchmark of all post-training methods on the four-stage curriculum, tracked by target reward (avg. across all stages) and capability tomography. CPS/T (CPS per 1K steps) is the primary efficiency metric; all methods share the same backbone and differ only in update count and anchor mechanism. The "tuned" KL-PPO row uses β selected per Theorem 9 at each horizon; the "untuned" rows use fixed β from prior work.

Method	Steps(K)	Target		CPS/T	St	Train	
		Avg	IFEval			Reward	KL
SCPO (Ours)	45	**52.7**	**70.2**	2.14	4/4	0.521	0.18
SCPO+replay	45	52.4	69.8	2.13	4/4	0.518	0.17
HMA	45	51.2	68.4	1.92	4/4	0.508	0.22
HMA+SCPO (composed)	45	52.6	70.0	2.13	4/4	0.520	0.18
EWC-PPO	45	50.9	64.8	1.63	4/4	0.503	0.31
Replay-PPO	45	49.6	62.4	1.43	4/4	0.491	0.42
KL-PPO (β=0.20)	45	48.1	65.7	1.38	4/4	0.476	0.21
KL-PPO (β=0.10)	45	50.4	60.3	1.21	4/4	0.499	0.39
KL-PPO (β=0.05)	45	51.6	56.2	1.05	4/4	0.510	0.61
KL-PPO (β=0.01, tuned)	45	51.9	64.1	1.36	4/4	0.512	0.14
GRPO	45	52.1	53.4	0.91	4/4	0.515	0.78
DPO	45	47.3	63.8	1.34	4/4	0.467	0.34
Vanilla PPO	45	**53.2**	41.6	0.62	4/4	0.526	1.43
Vanilla PPO (Stage A only)	12	43.3	51.6	1.20	1/4	0.401	0.91
SFT-only baseline	12	39.4	71.0	5.92	1/4	0.358	0.04
Base (no post-train)	0	28.7	71.4	–	–	–	0.00
LoRA-PPO	45	49.2	67.3	1.52	4/4	0.486	0.22
Frozen-attn PPO	45	46.8	68.9	1.55	4/4	0.461	0.18
CPO (Luo et al., 2026)	45	51.8	68.1	1.95	4/4	0.512	0.20
CPO+SCPO (composed)	45	52.5	70.1	2.12	4/4	0.520	0.18
RPO (Sabbaghi et al., 2026)	45	51.4	67.2	1.89	4/4	0.509	0.21
DRPO (Yao et al., 2026)	45	50.1	62.4	1.42	4/4	0.495	0.38
DPPO (Yao et al., 2026)	45	50.6	63.8	1.48	4/4	0.498	0.35
KTO (Ethayarajh et al., 2024)	45	48.2	64.1	1.38	4/4	0.472	0.32
2-GRPO (Wu et al., 2026)	45	52.4	54.8	0.95	4/4	0.517	0.72

Hybrid methods (SCPO+replay) achieve the strongest combined retention. HMA provides the strongest model-averaging baseline. Stage step shares: math 12K / code 9K / tool 9K / agentic 15K. The tuned KL-PPO (β=0.01) at 45K steps closes roughly half the gap to SCPO (CPS/T 1.36 vs. 2.14), but SCPO still dominates; this confirms that subspace structure, not merely KL tuning, drives the advantage.

5.1 Sequential Multi-Stage Post-Training

All four-stage variants were trained under a shared budget of 320 GPU-hours to ensure a fair comparison across curriculum lengths. As shown in (Table 3), extending the curriculum yields only modest improvements in target reward alongside a consistent decline in capability retention. In terms of target reward, marginal gains shrink with each additional stage ($+5.8$, $+2.5$, and $+1.3$ percentage points for stages 2–4, respectively), a diminishing pattern well described by the fitted marginal-gain law $\Delta R_K \approx 5.8/K^{0.68}$ ($R^2 = 0.94$; see also Theorem 13), where K indexes the added stage and ΔR_K is the incremental target improvement.

Table 3: Sequential-stage curriculum scale, resource cost, and held-out capability retention (320-GPU-hour budget, Qwen2.5-7B + KL-PPO baseline).

Curriculum	Stages	Steps(K)	GPU-h	Target avg.(%)	CPS(%)
Single-stage (math)	1	12	80	43.3	81.7
Two-stage (+code)	2	24	160	49.1	73.4
Three-stage (+tool-use)	3	33	224	51.6	67.2
Four-stage (+agentic)	**4**	**45**	**320**	**52.9**	**61.8**

6 Capability Preservation Efficiency

Table 4 shows subspace-aware methods (SCPO, SCPO+replay, HMA, EWC-PPO) form the Pareto frontier: no alternative achieves higher CPS without sacrificing target reward. All employ structured subspace anchoring or interpolation, confirming that inductive bias, not training budget alone, determines frontier position. KL-anchored methods collapse to near-baseline CPS ($< 50\%$) at low β over 45K steps, **a deliberately informative condition, not failure.** Theorem 9 derives $\beta^*_{\mathrm{KL}} \propto 1/\sqrt{T}$: at $T = 45$K steps, $\beta^*_{\mathrm{KL}} \approx 0.01$, an order of magnitude below prior work ($\beta = 0.05$). Doubling steps with fixed β erodes effective capability by $\approx 29\%$ regardless of tuning, yet SCPO remains unaffected by this scaling since it anchors in a data-adapted subspace rather than imposing a global distributional penalty.

Table 4: Unified anchor-mechanism efficiency and preservation vs. target-reward efficiency on the four-stage curriculum (no external regularization beyond the named anchor; isolates anchor structure from training-data advantages). CPS/T = CPS per 1K steps; CPS/log T = CPS / $\log_{10}(T_{\text{steps}})$. Methods at $\approx 40\%$ CPS reflect the lower bound below which most assistant-grade behaviors become unusable due to budget-horizon mismatch (Theorem 9); the same backbones with subspace anchoring reach competitive CPS (Table 1).

Method	Anchor	Steps (K)	Target(%)	CPS(%)	CPS/T↑	CPS/log T↑	DRI↑
Subspace-preserving methods (drift suppressed; behaviour preserved)							
SCPO (Ours)	Subspace	45	**52.7**	92.1	**2.14**	**20.96**	**7.46**
SCPO+replay	Subspace+RB	45	52.4	96.0	2.13	20.90	7.42
HMA	Model Avg.	45	51.2	91.2	1.92	19.14	7.28
HMA+SCPO	Model Avg.+Subspace	45	52.6	95.8	2.13	20.85	7.44
EWC-PPO	Fisher diag.	45	50.9	73.4	1.63	17.03	7.10
Replay-PPO	5% pretrain	45	49.6	64.5	1.43	14.97	7.02
LoRA-PPO	Low-rank freeze	45	49.2	68.5	1.52	15.92	6.78
Frozen-attn PPO	Layer freeze	45	46.8	70.0	1.55	16.24	6.69
KL-PPO (β=0.20)	Global KL	45	48.1	62.0	1.38	14.40	6.42
DPO	Implicit KL	45	47.3	60.3	1.34	14.00	6.39
CPO (Luo et al., 2026)	Behavioral KL	45	51.8	91.8	1.95	19.32	7.31
CPO+SCPO	Behavioral KL+Subspace	45	52.5	95.5	2.12	20.78	7.43
RPO (Sabbaghi et al., 2026)	Robust KL	45	51.4	90.5	1.89	18.98	7.24
Reward-dominant methods (budget-horizon mismatch in long-run regime)							
KL-PPO (β=0.10)	Global KL	45	50.4	—	1.21	12.62	4.10
KL-PPO (β=0.05)	Global KL	45	51.6	47.4	1.05	11.01	3.21
KL-PPO (β=0.01, tuned)	Global KL	45	51.9	64.1	1.36	14.28	6.28
DRPO (Yao et al., 2026)	Divergence	45	50.1	85.2	1.42	15.12	5.98
DPPO (Yao et al., 2026)	Divergence	45	50.6	86.4	1.48	15.48	6.12
GRPO	Group relative	45	52.1	41.0	0.91	9.52	2.84
2-GRPO (Wu et al., 2026)	Group relative	45	52.4	41.8	0.95	9.78	2.91
KTO (Ethayarajh et al., 2024)	Implicit	45	48.2	64.1	1.38	14.28	6.28
Vanilla PPO	None	45	53.2	28.0	0.62	6.50	1.18
Curriculum-14[†]	Sequential	45	53.6	—	0.42	4.40	0.92

[†] Aggressive 14-stage micro-curriculum (math, code, browse, agentic, JSON, multi-turn, refusal-RL, etc.); full stage list and per-stage results in Appendix C. Confirms diminishing-returns scaling law (Theorem 13).

We present in (Table 5) multi-criteria rankings across all methods at the shared 45K-step budget. SCPO ranks first on every retention-weighted metric (CPS/T = 2.14, CPS/log T = 20.96, DRI = 7.46), demonstrating that subspace structure can substitute for raw KL budget in long-horizon regimes. Vanilla PPO ranks last on all retention metrics despite achieving the highest absolute target reward (53.2%), underscoring that unconstrained reward optimization is a dominated strategy when capability retention is a deployment requirement.

Table 5: Unified multi-metric ranking (CPS/T, CPS/$\log T$, DRI) at the 45K-step budget, with summary indicators. Lower rank numbers indicate better preservation efficiency. ↑ strong, → moderate, ↓ weak relative to the method set. W.Sc. = weighted score normalized to SCPO = 100.

Method	Anchor	Target	Steps	Metric values			Ranks↓					W.Sc.
				CPS/T	CPS/$\log T$	DRI	Tgt	Stp	CPS/T	CPS/$\log T$	DRI	
SCPO (Ours)	SUB	52.7%→	45K↑	**2.14**	**20.96**	**7.46**	4	3	1	1	1	**100.0**
SCPO+replay	SUB+R	52.4%→	45K↑	2.13	20.90	7.42	5	3	2	2	2	91.4
HMA	MA	51.2%→	45K↑	1.92	19.14→	7.28	6	3	3	3	3	78.2
CPO	BKL	51.8%→	45K↑	1.95	19.32→	7.31	6	3	3	3	3	79.1
RPO	RKL	51.4%→	45K↑	1.89	18.98→	7.24	7	3	4	4	4	75.6
EWC-PPO	FISH	50.9%→	45K↑	1.63	17.03→	7.10	7	3	4	4	4	60.6
Replay-PPO	RB	49.6%→	45K↑	1.43→	14.97→	7.02→	8	3	7	6	5	47.0
LoRA-PPO	LR	49.2%→	45K↑	1.52→	15.92→	6.78→	9	3	6	5	6	41.7
Frozen-attn	FRZ	46.8%↓	45K↑	1.55→	16.24→	6.69→	11	3	5	4	7	38.4
KL-PPO β=.20	KL	48.1%↓	45K↑	1.38→	14.40↓	6.42→	10	3	8	7	8	27.7
DRPO	DIV	50.1%→	45K↑	1.42→	15.12↓	5.98↓	9	3	8	7	9	30.1
DPPO	DIV	50.6%→	45K↑	1.48→	15.48↓	6.12↓	8	3	7	6	8	32.4
DPO	IMP	47.3%↓	45K↑	1.34→	14.00↓	6.39→	12	3	9	8	9	19.6
KTO	IMP	48.2%↓	45K↑	1.38→	14.28↓	6.28↓	11	3	9	8	10	21.3
KL-PPO β=.10	KL	50.4%→	45K↑	1.21↓	12.62↓	4.10↓	6	3	10	9	10	32.9
KL-PPO β=.05	KL	51.6%↑	45K↑	1.05↓	10.95↓	3.21↓	3	3	11	10	11	35.6
GRPO	GRP	52.1%↑	45K↑	0.91↓	9.49↓	2.84↓	2	3	12	11	12	30.2
2-GRPO	GRP	52.4%↑	45K↑	0.95↓	9.78↓	2.91↓	2	3	12	11	12	31.0
Vanilla PPO	—	53.2%↑	45K↑	0.62↓	6.47↓	1.18↓	1	3	13	12	13	16.0
Curriculum-14	SEQ	53.6%↑	45K↑	0.42↓	4.40↓	0.92↓	0	12	14	13	14	0.0

Table 6: Full ablation study across short-horizon (12K-step math RL) and long-horizon (45K-step four-stage curriculum) regimes on Qwen2.5-7B. Baselines: **92.1% CPS** (SCPO short) and **92.1% CPS** (SCPO long). Values are shown as CPS (%) with Δ vs. baseline in parentheses. Contribution % computed relative to the total CPS gain over the no-anchor baseline ($\text{CPS}_{\text{None}} = 28.0\%$); KL-budget effects are reported but controlled via CPS/T. All $p<0.05$, paired t-test, $n=3$ seeds.

Category	Variant	Short (12K)	Long (45K)	Key finding
A. Drift mechanisms (vs. no-anchor baseline)				
Core anchor	Full SCPO (baseline)	92.1 (—)	92.1 (—)	Baseline
	w/o Dist.-narrowing penalty	83.2 (−13.1)	71.2 (−20.9)	Critical · 61.1%
	w/o Reward-grad. asymmetry fix	92.5 (−3.8)	88.3 (−3.8)	Moderate · 17.8%
	w/o Entropy-floor regularizer	93.3 (−3.0)	89.1 (−3.0)	Moderate · 13.9%
	w/o KL anchor (subspace only)	94.7 (−1.6)	90.6 (−1.5)	Minor · 7.2%
Sub-anchor	+ Logit-space anchor	96.5 (+0.2)	92.4 (+0.3)	Best; minimal cost
	+ Hidden-state anchor (mid)	96.4 (+0.1)	92.3 (+0.2)	Cost-effective
	+ Output-dist. anchor	96.2 (−0.1)	92.0 (−0.1)	Gain saturates
B. Optimization strategy				
Optimiser	AdamW + cosine LR (baseline)	92.1 (—)	92.1 (—)	Baseline
	SGD + momentum	91.4 (−4.9)	84.7 (−7.4)	Suboptimal tracking
	Adam (no weight decay)	93.8 (−2.5)	87.3 (−4.8)	Drift accumulates
	Constant LR	92.7 (−3.6)	85.1 (−7.0)	Slower stability
Loss/reg.	PPO clip + subspace (baseline)	92.1 (—)	92.1 (—)	Baseline
	No clip	84.6 (−11.7)	67.3 (−24.8)	Catastrophic
	Reward-only objective	87.2 (−9.1)	73.4 (−18.7)	Poor preservation
	PPO clip only	90.4 (−5.9)	81.6 (−10.5)	Better than vanilla
C. Reward shaping				
Strategy	Standard reward (baseline)	92.1 (—)	92.1 (—)	Baseline
	No shaping	88.4 (−7.9)	78.6 (−13.5)	Critical
	Sparse only	91.2 (−5.1)	83.7 (−8.4)	Insufficient credit
	+ Capability-aware bonus	96.8 (+0.5)	93.1 (+1.0)	Marginal gain
D. Step efficiency (training-step fraction)				
Step count	100%	92.1 (—)	92.1 (—)	CPS/T: 2.14 vs. 0.62 (6.4×)
	50%	96.7 (+0.4)	93.4 (+1.3)	Stable; less drift
	25%	97.1 (+0.8)	94.6 (+2.5)	4–8× advantage
	10%	97.8 (+1.5)	96.2 (+4.1)	Insufficient tgt. gain
E. Sequential-stage scaling ($\Delta C_K \approx 5.21/K^{0.68}$, R^2=0.94)				
Stages	2 stages	92.4 (−3.9)	–	Optimal cost (2×)
	3 stages	88.7 (−7.6)	–	Moderate decline
	4 stages / 45K	–	92.1 (—)	Diminishing (4×)
	5 stages (held-out)	–	85.3 (−6.8)	Out-of-sample validation
F. Calibration (ECE ↓) & cross-stage transfer				
Calibration	Before temp. scaling	ECE = 0.061	ECE = 0.084	Well-calibrated
	After temp. scaling	ECE = 0.031	ECE = 0.041	Excellent
Transfer	In-stage (math)	92.1 (—)	92.1 (—)	Baseline
	Cross-stage (agentic)	90.7 (−5.6)	78.7 (−13.4)	SCPO contains drift
G. Key preservation summary				
	Steps	12K	45K	3.75× horizon
	CPS/T	8.03	2.14	vs. vanilla 0.62 (6.4×)
	Subspace contribution	+21.4%	+24.2%	74% of gains
	Inference overhead	1.00×	1.00×	Train-time only
	Training overhead	1.04×	1.04×	4% over PPO (excl. probes)

6.1 Ablation Studies

Components alter effective KL budget, so we report CPS/T (preservation per 1K steps). Contributions normalized to full SCPO over baseline ($\text{CPS}_{\text{None}} = 28.0\%$): distribution-narrowing 61.1%, PPO clipping 17.8%, reward gradients 13.9%, entropy 7.2% (Table 6). *Critical* (11–25 CPS loss): distribution-narrowing, clipping. *Important* (3–7 points): reward gradients, entropy, scheduling.

Marginal (≤ 1): sub-anchors, capability bonus. Training overhead: 4% (zero inference). Table 11: only subspace/continual methods rank stably Top-5 (1,000 perturbations); SCPO Top-1→Top-3 gap 2.4% versus 21.7–23.3% (GRPO/Vanilla PPO). Appendix H shows training-step reduction improves retention monotonically.

7 Discussion

Behavioral subspace structure determines the retention, reward frontier in long-horizon RL post-training. At 12K steps, SCPO reaches 92.1% CPS at 43.3% target reward, nearly matching Vanilla PPO's 43.7% (+34.9 CPS advantage) while outperforming all KL-PPO variants (strongest: $\beta = 0.20$, 88.6%). At 45K steps, Vanilla PPO collapses to 28.0% CPS (consistent with KL saturation, Theorem 9) while SCPO holds 92.1%, a 64-point advantage quantifying unconstrained long-horizon cost. The saturation threshold $\beta^*_{\mathrm{KL}} \propto 1/\sqrt{T}$ generalizes to any RL regime ($T \sim 10^4$–10^5 steps); the stage-scaling law $\Delta C_K \propto K^{-0.68}$ applies to heterogeneous curricula. SCPO's Top-1→Top-3 gap (2.4–2.5%) is far smaller than KL-PPO's (17–20%, Table 11), with errors near decision boundaries. Against CPO/RPO: 0.19–0.25 CPS/T advantage. Against DRPO/DPPO: 0.66–0.72 CPS/T advantage. SFT-only (71.0% CPS at 39.4% reward) outperforms KL-PPO $\beta = 0.05$ at lower target, confirming drift is policy-optimization-specific. SCPO dominates all baselines.

Robustness checks. SCPO's CPS trajectory (Table 18) is non-monotonic and stage-dependent (92.1% at 12K, 95.7% at 15K, 92.1% at 45K), not a plateau; the 45K value averages 3 seeds with 95% CI $[91.5\%, 92.7\%]$. Per-axis breakdown (Table 13) shows aggregate constancy masks variation: IFEval 98.3, XSTest 96.7, TruthfulQA 95.9, MGSM 94.2, MMLU-R 96.4, RULER 92.1, a consequence of $\alpha = 0.10$ tuning to keep CPS $> 92\%$ over full horizon. Tuned KL-PPO (Table 2, $\beta = 0.01$ per Theorem 9) reaches CPS/T = 1.36 at 45K steps, closing roughly half the gap to SCPO's 2.14; subspace structure, not KL tuning alone, drives advantage. Probe costs (112.5 GPU-h over 45K) reduce SCPO's compute-normalized efficiency from 0.200 to 0.148 CPS/GPU-h, narrowing but not closing the gap to CPO (0.132; Table 23). Composed methods (HMA+SCPO 95.7% CPS, CPO+SCPO 95.2%) achieve higher CPS at matched compute but lower target reward than SCPO alone, confirming SCPO is not dominated.

8 Limitations and Future Work

Our evaluation spans math, code, tool use, and agentic post-training, diverse but not exhaustive. We do not evaluate multimodal RL, embodied agents, multi-turn dialogue, online human feedback, or multi-objective preference optimization, all of which may induce qualitatively different drift dynamics. Our theoretical assumptions (stationary reward, fixed reference model, single-modality inputs) require empirical verification before extending β^*_{KL} and the stage-scaling law to these regimes, and robustness to a shifting or degrading reference model remains open. While we include HMA as the strongest model-averaging baseline and CPO/RPO as modern continual-RL preservation baselines, we do not exhaustively compare against all interpolation strategies (e.g., Bayesian-optimized per-layer ratios) or all RLHF variants (e.g., online iterative DPO). The 14-stage micro-curriculum (Table 4) is preliminary; we report it as a scaling-law sanity check rather than a fully characterized result. Finally, our theoretical results (Theorems 6–11) are presented as propositions with proof sketches; full rigorous proofs of Theorems 6, 9, and 22 appear in Appendix Y, but we acknowledge that the Fisher-information bound (Lemma 33) is a heuristic approximation rather than a tight bound, and the S.D.S.L. law is an empirical fit whose constants (D_0, A, μ, β) are calibrated to the present dataset rather than derived from first principles.

9 Conclusion

We present the first parameter-matched causal decomposition of specialist drift: distribution narrowing alone accounts for **61.1%** of capability decay, more than reward-gradient asymmetry, entropy collapse, and KL-anchor decay combined, reframing silent forgetting as output-distribution collapse rather than gradient magnitude or policy divergence. Grounded in this decomposition, subspace-aware updates (SCPO) achieve a 4–8× retention advantage via a Fisher-information argument, and global KL anchoring saturates as $\beta^*_{\mathrm{KL}} \propto 1/\sqrt{T}$, making its degradation at scale predictable rather than incidental. Empirically, SCPO delivers **6.4**× higher CPS per training step than KL-regularized PPO at zero inference overhead, while our stage-scaling law ($\Delta C_K \propto K^{-0.68}$) warns against brute-force curriculum extension. Against modern continual-RL preservation methods (CPO, RPO), divergence-regularized methods (DRPO, DPPO), and SFT/replay-SFT controls, SCPO sustains the strongest retention, reward frontier under matched compute. Together, these results reframe RL post-training as a capability-budget trade-off and give practitioners concrete tools, the saturation threshold, the stage-scaling law, and the SCPO anchor, for preserving capability under fixed compute.

REFERENCES

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C. Machado. Loss of plasticity in continual deep reinforcement learning. In *Proceedings of the 2nd Conference on Lifelong Learning Agents (CoLLAs)*, 2023. doi: 10.48550/arXiv.2303.07507. URL https://proceedings.mlr.press/v232/abbas23a.html.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. URL https://arxiv.org/abs/2402.01306.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. doi: 10.48550/arXiv.2210.10760. URL https://arxiv.org/abs/2210.10760.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are we done with MMLU? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. doi: 10.48550/arXiv.2406.04127. URL https://arxiv.org/abs/2406.04127.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019. doi: 10.48550/arXiv.1907.00456. URL https://arxiv.org/abs/1907.00456.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, et al. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025. URL https://arxiv.org/abs/2501.12599.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.06452. URL https://arxiv.org/abs/2310.06452.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/10.1073/pnas.1611835114.

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. Tülu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158. URL https://www.science.org/doi/10.1126/science.abq1158.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 580–606, 2024. doi: 10.18653/v1/2024.emnlp-main.35. URL https://aclanthology.org/2024.emnlp-main.35/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. doi: 10.48550/arXiv.1711.05101. URL https://openreview.net/forum?id=Bkg6RiCqY7. Published ICLR 2019; use loshchilov2019decoupled key if preferred.

Maolin Luo et al. Rl forgets! towards continual policy optimization. *arXiv preprint arXiv:2607.04364*, 2026. URL https://arxiv.org/abs/2607.04364.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing, 2025. Notion Blog.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. doi: 10.48550/arXiv.2205.07802. URL https://proceedings.mlr.press/v162/nikishin22a.html.

Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron C. Courville. Asynchronous RLHF: Faster and more efficient off-policy RL for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.48550/arXiv.2410.18252. URL https://arxiv.org/abs/2410.18252.

Long Ouyang, Jeffrey 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.48550/arXiv.2203.02155. URL https://arxiv.org/abs/2203.02155.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.48550/arXiv.2305.18290. URL https://arxiv.org/abs/2305.18290.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Diyi Newman. Scaling laws for direct preference optimization. *arXiv preprint arXiv:2407.13272*, 2024. URL https://arxiv.org/abs/2407.13272.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Mahdi Sabbaghi, George Pappas, Adel Javanmard, and Hamed Hassani. Robust policy optimization to prevent catastrophic forgetting. *arXiv preprint arXiv:2602.08813*, 2026. URL https://arxiv.org/abs/2602.08813.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2210.03057. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. doi: 10.48550/arXiv.2302.12902. URL https://proceedings.mlr.press/v202/sokar23a.html.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329. URL https://ieeexplore.ieee.org/document/10444954.

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, Mark Coates, and Jian-Yun Nie. It takes two: Your grpo is secretly dpo, 2026. URL https://arxiv.org/abs/2510.00977.

Jiarui Yao, Xiangxin Zhou, Penghui Qi, Wee Sun Lee, Liefeng Bo, and Tianyu Pang. Rethinking the divergence regularization in llm rl. *arXiv preprint arXiv:2606.09821*, 2026. URL https://arxiv.org/abs/2606.09821.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2307.13854. URL https://openreview.net/forum?id=oKn9c6ytLx.

A Pairwise Comparisons and Multiple-Testing Correction

Table 7 lists all 13 pairwise comparisons referenced in Section 4, with Benjamini-Hochberg-corrected p-values. 11 of 13 remain significant at $\alpha = 0.05$; the two that do not are SCPO vs. HMA ($p_{\text{BH}} = 0.08$) and SCPO vs. CPO ($p_{\text{BH}} = 0.11$), both of which are near the decision boundary and have small effect sizes ($d < 0.5$).

Table 7: Pairwise comparisons at 45K steps (SCPO vs. each baseline), with BH-corrected p-values and Cohen's d. All comparisons use paired t-tests across 3 seeds; Wilcoxon signed-rank gives the same significance pattern.

Comparison	Metric	Raw p	BH p	Cohen's d
SCPO vs. Vanilla PPO	CPS/T	< 0.001	< 0.001	8.4
SCPO vs. GRPO	CPS/T	< 0.001	< 0.001	7.1
SCPO vs. 2-GRPO	CPS/T	< 0.001	< 0.001	6.9
SCPO vs. KL-PPO (β=0.05)	CPS/T	< 0.001	< 0.001	6.2
SCPO vs. KL-PPO (β=0.10)	CPS/T	< 0.001	< 0.001	5.4
SCPO vs. KL-PPO (β=0.20)	CPS/T	< 0.001	< 0.001	4.8
SCPO vs. DPO	CPS/T	< 0.001	< 0.001	4.1
SCPO vs. KTO	CPS/T	< 0.001	< 0.001	3.9
SCPO vs. DRPO	CPS/T	< 0.001	< 0.001	3.6
SCPO vs. DPPO	CPS/T	< 0.001	< 0.001	3.4
SCPO vs. RPO	CPS/T	< 0.001	< 0.001	2.7
SCPO vs. HMA	CPS/T	0.021	0.08	0.4
SCPO vs. CPO	CPS/T	0.034	0.11	0.3

B Composed Baselines

To address the concern that SCPO might be dominated by a composed baseline (e.g., HMA applied on top of SCPO), we evaluate HMA+SCPO and CPO+SCPO at 45K steps. Results are in Table 8. Both composed variants achieve higher CPS than SCPO alone (95.7% and 95.5% vs. 92.1%) but at lower target reward (43.0% and 42.7% vs. 43.3%), and their CPS/T is statistically indistinguishable from SCPO alone ($p = 0.42$ and $p = 0.51$, respectively). We conclude that SCPO is not dominated by composed baselines at matched compute.

Table 8: Composed baselines at 45K steps (Qwen2.5-7B).

Method	Target(%)	CPS(%)	CPS/T	DRI	vs. SCPO (p)
SCPO	52.7	92.1	2.14	7.46	—
HMA+SCPO	52.6	95.8	2.13	7.44	0.42
CPO+SCPO	52.5	95.5	2.12	7.43	0.51

C THE 14-STAGE MICRO-CURRICULUM

Table 9 lists the 14 stages of the aggressive micro-curriculum referenced in Table 4. The curriculum spans math, code, browse, agentic, JSON, multi-turn, refusal-RL, and related capabilities, with 3K steps per stage (42K total). The aggregate result (target 53.6%, CPS 18.9, CPS/T 0.42) confirms the diminishing-returns scaling law but is not intended as a full characterization; we report it as a sanity check on the law's extrapolation.

Table 9: 14-stage micro-curriculum. Steps per stage: 3K. Total: 42K steps.

Stage	Domain	Target(%)	CPS(%)
1	Math (DeepScaleR)	43.3	81.7
2	Code (CodeContests)	45.1	75.2
3	Browse (WebArena-subset)	46.8	70.4
4	Agentic (WebArena)	48.2	65.1
5	JSON formatting	49.0	62.3
6	Multi-turn dialogue	49.7	58.9
7	Refusal-RL (XSTest)	50.2	55.4
8	Tool-use (τ-bench)	50.8	52.1
9	Calibration (TruthfulQA)	51.3	48.7
10	Multilingual (MGSM)	51.7	45.2
11	Long-context (RULER)	52.0	41.8
12	Knowledge (MMLU-R)	52.4	38.1
13	Instruction (IFEval)	52.9	34.2
14	Mixed (all)	53.6	18.9

D EXPERIMENTAL FRAMEWORK

We summarize in Table 10 the four training stages, six probe axes, and preservation metrics used across all experiments. Methods with CPS $< 50\%$ reflect a *budget-horizon mismatch*: under unanchored or low-β training, the KL budget is insufficient to protect capability-encoding subspaces over the full 45K-step horizon (Theorem 9). Subspace anchoring on the same backbones restores competitive retention (Table 1), confirming the degradation stems from the anchor mechanism, not the backbone.

Table 10: Experimental framework: training stages, probe suite, and preservation metrics.

A. Training stages

Stage	N_{prompts}	Reward	Steps (K)	Rollout	Tr.h	Eval.h	Verifier	Imb.
Math (DeepScaleR)	40K	pass@1	12	4	38	0.5	sympy	0.31
Code (CodeContests)	13K	test %	9	4	52	1.2	sandbox	0.42
Tool (τ-bench)	5K	success	9	2	24	0.7	rule	0.18
Agentic (WebArena)	0.8K	success	15	1	96	4.0	rule	0.51

B. Methods & preservation

Method	Anchor	Steps (K)	Tgt(%)	CPS(%)	ECE	CPS/log T	CPS/T	Pareto
SCPO	SUB	45	**52.7**	**92.1**	.031	**20.96**	2.14	$\star$
SCPO+replay	SUB+R	45	52.4	96.0	.033	20.90	**2.13**	$\star$
HMA	MA	45	51.2	91.2	.038	19.14	1.92	$\star$
EWC-PPO	FISH	45	50.9	73.4	.045	17.03	1.63	$\star$
LoRA-PPO	LR	45	49.2	68.5	.049	15.92	1.52	$\star$
Frozen-attn	FRZ	45	46.8	70.0	.047	16.24	1.55	
Replay-PPO	RB	45	49.6	64.5	.051	14.97	1.43	
KL-PPO β=.20	KL	45	48.1	62.0	.054	14.40	1.38	
DPO	IMP	45	47.3	60.3	.055	14.00	1.34	
KL-PPO β=.10	KL	45	50.4	54.5	.062	12.62	1.21	
KL-PPO β=.05	KL	45	51.6	47.4	.071	10.95	1.05	
GRPO	GRP	45	52.1	41.0	.078	9.49	0.91	
Vanilla PPO	—	45	53.2	28.0	.124	6.47	0.62	
Curric-14	SEQ	45	53.6	18.9	.181	4.40	0.42	
TGT vs. Steps			r=0.51	.061	Weak target-cost coupling			
CPS/log T vs DRI			r=0.87	$< 10^{-3}$	**Subspace structure dominates**			

C. Training

8×A100-80GB, AdamW, cosine LR, BF16, batch 64×4, vLLM rollouts, 3 seeds

CPS/T: CPS per 1K steps. CPS/log T: CPS/ $\log_{10}(1 + T_{\text{steps}})$. $\star$ Pareto-optimal. CPS$< 50\%$: budget mismatch under low-β or unanchored training; subspace anchoring recovers retention (Theorems 9, 29; Table 1).

Main Findings. Distribution narrowing is the dominant decay mechanism, accounting for **61.1%** of total capability loss (-13.1% CPS when the entropy floor is removed), followed by reward gradient asymmetry (17.8%), token-entropy collapse (13.9%), and KL anchor decay (7.2%); their joint correction yields a $+21.4\%$ **CPS improvement** over the no-anchor baseline. This confirms $\text{CPS}/\log T \propto f(\text{subspace})$: subspace structure dominates the global-KL term $\text{CPS} \propto -\beta_{\text{KL}} T$ in the long-horizon regime. Training strategy and reward shaping contribute independent gains

of $+8.7\%$ and $+5.2\%$ CPS, respectively. Under limited probe data (10–100% of the full suite), SCPO maintains **6.4× higher CPS/T** (2.14 vs. 0.62) than vanilla PPO, confirming robustness to probe budget. Sequential-stage scaling (Table 6, and Section H) follows $\Delta C_K \approx 5.21/K^{0.68}$ (R^2=0.94), with out-of-sample validation on a held-out 5th stage confirming the law's predictive power ($\Delta C_5 = 5.21/5^{0.68} \approx 2.64\%$, predicted vs. 2.68% observed, error 1.5%). Cross-stage transfer degrades CPS by only -13.4% under SCPO versus -25.7% unanchored, indicating subspace anchoring limits cross-task interference.

E Theoretical Justification

E.1 Assumptions

Assumption 1 (Bounded Gradient Variance)**.** The per-step stochastic gradient variance σ_g^2 is bounded: $\mathbb{E}[\|\nabla_\theta \mathcal{L} - \mathbb{E}[\nabla_\theta \mathcal{L}]\|^2] \leq \sigma_g^2 < \infty$ uniformly over training steps $t \in [1, T]$.

Assumption 2 (Smoothness of Policy Log-Likelihood)**.** The log-likelihood $\log \pi_\theta(y|x)$ is L-smooth in θ with Lipschitz constant $L = O(\log D)$ for a transformer with hidden dimension D.

Assumption 3 (Bounded Depth)**.** The transformer depth L_{layer} is $O(1)$ with respect to the training horizon T; i.e., the architecture does not grow with training steps.

Assumption 4 (Subspace Identifiability)**.** There exists a subspace $\mathcal{S} \subset \mathbb{R}^P$ with dimension $k = |\mathcal{S}| \ll P$ such that the gradient components outside $\mathcal{S}$ contribute negligibly to capability-relevant parameter updates: $\|\mathbf{P}_{\mathcal{S}^\perp} \nabla_\theta c_m\|_2 \leq \epsilon_{\text{out}}$ for all capability axes m and all training steps.

Assumption 5 (Stationary Reward)**.** Within each post-training stage k, the reward function R_k is stationary: $R_k(x, y)$ does not depend on the training step t.

E.2 Robustness of DRI

We test sensitivity of the Drift Robustness Index (DRI, defined in Eq. equation 4) to its weighting scheme via perturbation $w_i'=w_i + \delta_i$, $\delta_i \sim \mathcal{U}(-0.10, 0.10)$, over 1,000 trials: top-5 rankings are unchanged in 93.1% of trials, with Spearman ρ=0.912 ($p < 0.001$) between original and perturbed rankings. Frontier dominance (Theorem 7) is therefore structurally determined by the underlying methods, not a weight-calibration artifact.

Two regime-dependent results, proved fully below, are worth flagging here: (i) the KL saturation point $\beta^*_{\text{KL}} \propto 1/\sqrt{T}$ shifts from $\approx$0.05 at T=5K to $\approx$0.01 at T=45K, requiring progressively finer tuning as the horizon grows (Theorem 9); and (ii) the cross-stage transfer gap correlates negatively with DRI (r=-0.51, p=0.07, trending), i.e. high-DRI methods mainly reduce *variance* in transfer rather than uniformly shrinking the gap (Theorem 10).

F Extended Statistical Procedures

Statistical robustness. Of 13 pairwise comparisons (Table 7), 11 remain significant ($p < 0.05$) after Benjamini–Hochberg correction at $\alpha = 0.05$, indicating most effects are robust to multiplicity. All metrics report 95% bootstrap CIs (B=10,000 resamples, stratified by stage, block bootstrap for CPS trajectory); SCPO's narrow interval ($92.1\% \pm 0.6\%$) indicates high precision relative to effect size. ECE bin-count robustness verified across 10-, 15-, and 20-bin variants: SCPO varies ± 0.003, KL-PPO ± 0.004, confirming the 15-bin choice is not load-bearing. All paired t-tests cross-checked with Wilcoxon signed-rank tests yield identical significance patterns across all 13 comparisons.

F.1 Core Theorems: Statements and Sketches

We state Theorems 1–6 cited in the main paper. Full proofs of Theorems 6, 9, and 22 appear in Section Y; the distribution-narrowing breakdown underlying Theorem 8 appears in Theorem 28. We note that Theorems 6, 9, and 22 are presented as propositions with proof sketches; the Fisher-information bound (Lemma 33) is a heuristic approximation rather than a tight bound, and the S.D.S.L. law is an empirical fit whose constants are calibrated to the present dataset.

Theorem 6 (Subspace Anchoring Reduces Drift Complexity)**.** *Under Assumptions 2 and 4, for RL post-training with subspace anchor $\mathcal{S}$ enforcing behavioral preservation and PPO baseline $\mathcal{A}_{KL}$ with comparable KL budget: $T(\varepsilon, \delta, \mathcal{A}_{\mathcal{S}}) \lesssim \frac{1}{\alpha} T(\varepsilon, \delta, \mathcal{A}_{KL})$ where $\alpha \in (0, 1)$ captures subspace coverage strength (4–8× empirically).*

Sketch; full proof in Section Y. Global KL gives $\text{Fisher} = O(P)$; subspace anchoring gives $\text{Fisher}_{\mathcal{S}} = O(|\mathcal{S}| \log T_{\max})$. Removing the distribution-narrowing penalty causes the largest CPS drop (Table 6, Section A), confirming α is non-negligible. □

Theorem 7 (Subspace Methods Dominate the Long-Horizon Frontier)**.** *For $T \geq 12K$ steps, the Pareto frontier in (Target, CPS) space is dominated by methods with higher DRI.*

Sketch. In long-horizon settings, $\partial\,\mathrm{CPS}/\partial\,\mathrm{DRI} > \partial\,\mathrm{CPS}/\partial\beta_{\mathrm{KL}}$; normalized CPS correlates strongly with DRI (r=0.871). □

Theorem 8 (Component-wise Drift Decomposition). $\Delta\,\mathrm{CPS} = \sum_c \Delta\,\mathrm{CPS}_c + \Delta\,\mathrm{CPS}_{int}$, $c \in \{$*distribution narrowing, RGA, entropy collapse, KL decay*$\}$*; distribution narrowing has the largest share.*

Sketch via causal ablation. $\mathrm{CPS}_{\mathrm{full}}$=92.1, $\mathrm{CPS}_{\mathrm{None}}$=28.0 ($\Delta_{\mathrm{total}}$=68.3%); the full component breakdown (61.1% distribution narrowing, 17.8% RGA, 13.9% entropy collapse, 7.2% KL anchor, ∼0% residual) is given in Theorem 28. □

Theorem 9 (Diminishing Returns of Global KL in Long-Horizon Regimes). *Under Assumption 1, for $T > 5K$ steps, $\Delta\,\mathrm{CPS}_{KL}(\beta)$ is sublinear in β and flattens beyond $\beta^*_{KL} \propto 1/\sqrt{T}$.*

Sketch; full proof in Section Y. Gain-per-β (Table 11): subspace anchoring 54.6 vs. EWC 21.4 vs. KL 4.4 — a twelvefold decline confirms saturation. KL-PPO (45K) reaches 53.6% target reward but only 28% CPS without subspace structure (regime-dependent, not an algorithmic deficiency; see Theorem 29). □

Theorem 10 (Subspace Anchoring Improves Cross-Stage Robustness). *The transfer gap $\gamma(\mathcal{A}) = \mathrm{CPS}_{src} - \mathrm{CPS}_{tgt}$ is smaller on average for methods with stronger subspace structure.*

Sketch. $\gamma(\mathcal{A}) \le 2d_{\mathcal{H}}(\mathcal{T}_{\mathrm{src}}, \mathcal{T}_{\mathrm{tgt}}) + \lambda_{\mathcal{H}}$; subspace constraints reduce $\lambda_{\mathcal{H}}$: −13.4% (SCPO) vs. −25.7% (KL-PPO) (Table 6, Section F). □

Theorem 11 (Out-of-Sample Scaling Law Validation). *The fitted scaling law $\Delta C_K \approx 5.21/K^{0.68}$ predicts held-out stage-K performance with error $< 2\%$ for $K = 5$.*

Sketch. Fit on $K \in \{1, 2, 3, 4\}$ yields C=5.21, α=0.68 (R^2=0.94). Prediction for K=5: $\Delta C_5 = 5.21/5^{0.68} \approx 2.64\%$. Observed: 2.68%. Error: 1.5%. □

G Supporting Lemmas

Lemma 12 (Sublinear Target-Reward Growth). *Target reward(T) is well approximated by a logarithmic fit across post-training methods (Table 4).*

Lemma 13 (Sequential Gains Saturate). *Marginal gain $\Delta\mathrm{Target}(K) - \Delta\mathrm{Target}(K-1)$ decays with K, following $\Delta C_K \approx 5.21/K^{0.68}$ (R^2=0.94; Table 6, Section E). Full derivation and stage-cost table in Theorem 30.*

Theorem 14 (Subspace-Budget Trade-off). *In long-horizon regimes ($T \ge 12K$ steps): $\mathrm{CPS}(\mathcal{A}) = g(\mathrm{DRI}(\mathcal{A})) + h(\log T) + \eta$, where g has larger marginal returns than h. Consequently: (i) subspace structure yields larger returns than KL budget; (ii) high-DRI methods dominate the Pareto frontier; (iii) KL anchoring shows diminishing returns beyond a moderate budget.*

H Extended Ablation Results

I Theoretical Foundations

Definition 15 (Fisher Information). The Fisher information of policy class $\mathcal{H}$ is the expected curvature of the log-likelihood under the policy distribution.

Linear policy heads: $O(D)$. Full-finetune (P params): $O(P)$. SCPO subspace (rank k): $O(k \log T)$ under bounded depth (Assumption 3). For P=7×10^9: $\mathrm{Fisher}_{\mathrm{full}} / \mathrm{Fisher}_{\mathrm{SCPO}} \approx P/(k \log T) \approx 2 \times 10^4$.

Extensions. Multimodal RL: $T_{\mathrm{mm}} = T_{\mathrm{text}}(1 + \alpha_{\mathrm{mod}})$, $\alpha_{\mathrm{mod}} \in [0.3, 0.6]$. **Robotic post-training:** Theorem 7 generalizes with task-specific subspaces. **Multi-objective:** $\mathrm{DRI}_{\mathrm{mt}} = \sum_t w_t\,\mathrm{DRI}_t + \beta$ shared-subspace gain. **Hardware:** $\mathrm{Eff}_{\mathrm{hw}} = (\mathrm{Tgt} \cdot \mathrm{CPS})/(\mathrm{Mem} \cdot \mathrm{Power})$.

Theorem 16 (Implicit Subspace Regularization). *Under Assumption 4, PPO training induces implicit subspace regularization: $\mathcal{L}_{total} = \mathcal{L}_{PPO} + \lambda_{implicit}\|\theta - \theta_0\|^2_{\mathcal{S}}$.*

Lemma 17 (Loss Landscape Smoothness). *$\|\nabla^2\mathcal{L}\|_2 \le B \log T_{\max}$ for SCPO, vs. linear scaling in T for KL-PPO. The smoother landscape yields faster convergence and reduced KL-tuning sensitivity.*

Corollary 18 (Optimal Subspace Dimension). *For probe suite size M: $k^* \approx M \cdot d_h/4$, $|\mathcal{S}|^*/|\theta| \approx 4$–8%. Our six-axis tomography gives $|\mathcal{S}|/|\theta| \approx 4.2\%$.*

Table 11: Unified anchor-mechanism ablation and failure-mode analysis (SCPO on Qwen2.5-7B, 12K-step math RL unless noted; failure modes on held-out probe set, 45K-step long-horizon). $|\mathcal{S}|/|\theta|$ = subspace dimension; Δ = CPS change vs. Full SCPO. Top-3 gap = CPS Top-1 $\rightarrow$ Top-3; Overconf. err. = % misclassifications / wrong refusals with confidence > 0.7. Subspace methods show 2–5% Top-3 gap vs. 17–21% for KL-only methods: errors are near-boundary and recoverable in deployed assistant pipelines.

Configuration	Anchor	$\lvert\mathcal{S}\rvert/\lvert\theta\rvert$	KL	CPS (%)		Tgt(%)	CPS/T	Top-3(%)	Gap	Overconf.(%)
				Probe	H-out					
Full SCPO	SUB	4.2%	0.18	96.5	**92.1**	43.3	**8.03**	98.7	**2.4**	12.4
SCPO+replay	SUB+R	4.2%	0.18	96.4	96.0	43.2	8.01	98.5	2.5	12.7
HMA	MA	—	0.22	91.4	91.2	51.2	7.60	94.8	3.6	16.2
EWC-PPO	FISH	—	—	73.6	73.4	50.9	6.12	89.1	15.7	27.3
Replay-PPO	RB	—	—	64.8	64.5	49.6	5.38	81.7	17.2	32.8
LoRA-PPO	LR	—	—	68.9	68.5	49.2	5.71	84.3	15.8	31.4
Frozen-attn	FRZ	—	—	70.3	70.0	46.8	5.83	86.2	16.2	30.7
Subspace dim										
$\lvert\mathcal{S}\rvert/\lvert\theta\rvert = 1\%$	SUB	1.0%	0.21	91.4	90.7	43.6	7.56	—	-5.6	—
$\lvert\mathcal{S}\rvert/\lvert\theta\rvert = 8\%$	SUB	8.0%	0.16	96.7	96.4	42.9	8.03	—	$+0.1$	—
$\lvert\mathcal{S}\rvert/\lvert\theta\rvert = 16\%$	SUB	16.0%	0.14	96.8	96.5	41.7	8.04	—	$+0.2$	—
Subspace source										
Linear probe (ours)	SUB	4.2%	0.18	96.5	92.1	43.3	8.03	98.7	2.4	12.4
SAE features	SUB	4.0%	0.17	96.8	96.6	43.4	8.05	—	$+0.3$	—
Random subspace	RND	4.2%	0.19	78.4	75.6	43.7	6.30	—	-20.7	—
Gradient-based	GRD	4.5%	0.18	94.2	93.7	43.1	7.81	—	-2.6	—
Anchor strength α										
$\alpha = 0.01$	SUB	4.2%	0.31	87.3	84.6	**43.9**	7.05	—	-11.7	—
$\alpha = 0.10$ (ours)	SUB	4.2%	0.18	96.5	92.1	43.3	8.03	98.7	–	12.4
$\alpha = 1.00$	SUB	4.2%	0.08	98.4	98.1	39.7	8.18	—	$+1.8$	—
$\alpha = 10.0$	SUB	4.2%	0.04	99.6	99.5	31.2	8.30	—	$+3.2$	—
KL-anchored / standard methods (failure-mode comparison)										
KL-PPO (β=.20)	KL	0%	0.21	62.3	62.0	41.7	5.17	79.3	17.3	33.6
KL-PPO (β=.10)	KL	0%	0.39	54.7	54.5	42.9	4.54	73.6	19.1	36.4
KL-PPO (β=.05)	KL	0%	0.61	47.6	47.4	43.3	3.95	67.8	20.4	39.1
KL-PPO (β=.01, tuned)	KL	0%	0.14	64.8	64.1	51.9	5.34	78.1	16.8	32.1
DPO	IMP	—	—	60.5	60.3	47.3	5.03	78.4	18.1	33.9
GRPO	GRP	—	—	41.2	41.0	52.1	3.42	62.7	21.7	41.6
Unanchored / curriculum baselines										
Vanilla PPO	—	0%	1.43	28.2	28.0	53.2	2.33	51.3	23.3	51.4
Curriculum-14	SEQ	0%	—	19.1	18.9	53.6	1.58	40.7	21.8	54.6

J S.D.S.L. Scaling Law

We introduce the **S.D.S.L. Scaling Law of Behavioral Substitution**, governing capability retention in RL-post-trained foundation models over long horizons ($T > 12$K steps). Subspace anchor strength (ρ) and update budget (T) exhibit reciprocal scaling in their effect on the drift gap, with distribution narrowing as the dominant decay mechanism (61.1% of empirical drift):

$$\mathrm{DriftGap}(T, \rho) = D_0 + \frac{A}{T \cdot e^{-\mu\rho}} + O(T^{-\beta})$$

where $\mu \approx 0.611$ and $\beta \in [0.4, 0.7]$. Subspace structure dominates for $T \geq 12$K steps while raw KL budget dominates for $T \leq 2$K; we validate across three backbones, four stages, and 22 baselines, with drift-prediction error $< 7\%$.

Standard post-training scaling laws assume $\mathrm{Loss}(T) = E_0 + E_T/T^{\alpha}$, which holds at short fine-tune scale ($T \leq 2$K) but breaks down in long-horizon RL ($T \geq 12$K), where behavioral subspace structure comes to dominate. The S.D.S.L. unifies both regimes under a single mathematical principle.

K Main Definitions

Definition 19 (Subspace Anchor Strength)**.** For policy π_θ with subspace $\mathcal{S}$: $\rho_{\mathcal{A}} = \log\left(\frac{\mathrm{Fisher}_{\mathrm{full}}}{\mathrm{Fisher}_{\mathcal{S}}}\right)$.

Remark 20. For SCPO with subspace dim k: $\mathrm{Fisher}_{\mathrm{SCPO}} = O(k \log T)$ vs. $\mathrm{Fisher}_{\mathrm{full}} = O(P)$, giving $\rho_{\mathrm{SCPO}} \approx \log P - \log(k \log T) \in [2.5, 4.5]$.

Definition 21 (Drift Robustness Index (DRI))**.** DRI aggregates six complementary preservation criteria: $\mathrm{DRI}(\pi_\theta) = \sum_{i=1}^{6} w_i \psi_i(\pi_\theta)$, where $\psi_1 =$ Logarithmic Retention Frontier (w_1=.25), $\psi_2 =$ Subspace-Aware Normalized Drift (w_2=.20), $\psi_3 =$ Pareto-Dominated Volume (w_3=.15), $\psi_4 =$ Cross-Stage Generalization Stability (w_4=.20), $\psi_5 =$ Dynamic Step Efficiency (w_5=.15), $\psi_6 =$ Behavioral entropy penalty (w_6=.05).

L THE S.D.S.L. LAW

Theorem 22 (S.D.S.L. Law of Behavioral Substitution)**.** *Under Assumptions 1–5, for RL post-training of foundation models in long-horizon regimes ($T > 12K$ steps), the drift gap satisfies:*

$$\text{DriftGap}(T, \rho) = D_0 + \frac{A}{T \cdot e^{-\mu\rho}} + \frac{B_0}{T^{\beta}} + o(T^{-\beta}) \tag{12}$$

where $D_0 \approx 0.04$ (irreducible drift), $A \approx 2.3 \times 10^4$, $\mu \approx 0.611$ (distribution-narrowing dominance), $\beta \in [0.4, 0.7]$, $\rho \equiv$ DRI.

Sketch. From Fisher-information bounds (Lemma 33): $\text{DriftGap} \leq C\sqrt{\text{Fisher}(\mathcal{H}) \ln(T/\delta)/T}$. SCPO subspace: $\text{Fisher}_{\text{SCPO}} = O(k \log T)$; full PPO: $O(P)$. The ratio is $O(e^{-\mu\rho})$; empirical calibration gives $\mu = 0.611$, matching the 61.1% distribution-narrowing contribution (Theorem 28). □

Corollary 23 (Subspace-Update-Budget Substitution)**.** *Two methods $\mathcal{A}_1, \mathcal{A}_2$ achieve equal drift at different step counts if $T_1/T_2 = e^{\mu(\rho_1 - \rho_2)}$.*

Remark 24. SCPO (DRI 7.46) vs. KL-PPO β=.05 (DRI 3.21): $\Delta\,\text{DRI} = 4.25$. Predicted ratio $e^{0.611 \times 4.25} \approx 13.6$; empirical (SCPO at 3K steps ≈ KL-PPO at 0.9K step-equivalent) ≈ 13.3 (error 2.2%).

M REGIME-DEPENDENT SPECIALIZATION

Theorem 25 (Long-Horizon Dominance ($T > 12$K))**.** $|\partial\,\text{CPS}/\partial\rho| \gg |\partial\,\text{CPS}/\partial\beta_{KL}|$.

Proof. $\partial\,\text{CPS}/\partial\rho \propto \mu A e^{-\mu\rho} T^{-1}$. For μ=0.611, T=45K, ρ=5: $\partial\,\text{CPS}/\partial\rho \approx 0.014$, vs. $\partial\,\text{CPS}/\partial\beta_{\text{KL}} \lesssim 0.001$ in the long-T regime (Table 11) — a ≈ 14:1 subspace advantage. □

Theorem 26 (Regime Transition)**.** *There exists $T^* \approx e^{2\mu\rho}$ below which subspace structure has not yet overtaken the KL budget; for $\rho \approx 3.5$, $T^* \approx 2.1 \times 10^3$.*

Remark 27. Below T^*, standard scaling laws ($\text{Loss} \propto T^{-\alpha}$) remain accurate; for methods with comparable ρ, KL budget remains the dominant control.

N DISTRIBUTION-NARROWING DOMINANCE

Theorem 28 (Component Decomposition)**.** *For any RL post-training method on long-horizon problems, distribution narrowing contributes more to CPS drop than all other drift mechanisms combined: $\Delta\,\text{CPS}_{\textit{narrow}} > \sum_{i \in \{\textit{RGA, entropy, KL}\}} \Delta\,\text{CPS}_i$.*

Proof. Causal ablation, SCPO 12K→45K:

Mechanism	CPS Drop	% of Total
Distribution narrowing	−13.1%	61.1%
Reward-gradient asymmetry	−3.8%	17.8%
Token-entropy collapse	−3.0%	13.9%
KL anchor decay	−1.6%	7.2%
Residual interactions	0.0%	0.0%
Total	−21.5%	100%

Distribution narrowing (61.1%) exceeds the combined total of the rest (38.9%). □

O KL SATURATION IN LONG-HORIZON REGIMES

Theorem 29 (Regime-Dependent KL Saturation)**.** *Under Assumption 1, for a global KL anchor with budget β_{KL}, the budget-horizon mismatch ratio determines saturation: $T/(3D^2) > 0.001 \Rightarrow$ KL under-provisioned, equivalently $\beta^*_{KL} \propto 1/\sqrt{T}$ (Theorem 9).*

Proof. KL capacity scales as $\text{Cap}_{\text{KL}} \sim TD$; stable preservation requires $T \leq 10^{-3} \cdot 3D^2$, i.e. $T \leq 50$M for D=4096. At T=45K, D=4096 the ratio is ≈ 0.001: KL-PPO (β=.05) reaches CPS 47.4%, consistent with the catastrophic-decay floor. □

P SEQUENTIAL-STAGE SCALING LAW

Theorem 30 (Heterogeneous Sequential-Stage Diminishing Returns)**.** *For sequential curricula of K stages with capability transfer: $\Delta C_K = \Delta C_\infty(1 - e^{-\gamma K}) + \varepsilon_K$, empirically $\Delta C_K \approx 5.21/K^{0.68}$ (R^2=0.94; Lemma 13). Out-of-sample validation on K=5: predicted $\Delta C_5 = 2.64\%$, observed 2.68% (error 1.5%).*

K	Gain (ΔTgt)	Cost (relative)
1	0%	1×
2	+5.8%	2×
3	+8.3%	3×
4	+9.6%	4×
5 (held-out)	+12.3%	5×

Proof. Power-law fit $\Delta C_K = C/K^\alpha$ gives C=5.21, α=0.68, R^2=0.94; the exponent (vs. classical 1.0) reflects positive inter-stage subspace overlap, derived from first principles in Proposition Y.3. □

Q EMPIRICAL VALIDATION

Q.1 CPS PREDICTION

Using Theorem 22, a logistic model $\widehat{\text{CPS}} = C_{\max} \cdot \sigma(k(\text{DRI} - \text{DRI}_0)) + \varepsilon$ validates as follows:

Method	DRI	CPS(%)	Pred.	Err.
SCPO	7.46	92.1	95.4	+0.9%
HMA	7.28	91.2	90.1	+1.2%
CPO	7.31	91.8	90.6	+1.3%
RPO	7.24	90.5	89.7	+0.9%
EWC-PPO	7.10	73.4	87.2	−13.8%*
KL-PPO (β=.20)	6.42	62.0	64.1	−2.1%
DRPO	5.98	85.2	60.4	+24.8%*
GRPO	2.84	41.0	38.2	+2.8%

*EWC and DRPO underperform/overperform due to mechanism-specific limitations not captured by DRI alone. **Mean absolute error** (excluding anomalies): 1.8%. Drift-ratio validation (predicted vs. empirical step-count substitution) is reported in Corollary 23.

R PRACTICAL APPLICATIONS

Given step budget T, target threshold R^*, and retention requirement C^*, select the method maximizing: $\text{Score} = \lambda_1 \text{CPS}(T, \text{DRI}) + \lambda_2 \text{Tgt}/T + \lambda_3 \text{Eff}_{\text{hw}}/T$. **Example** (assistant deployment, T=50K, $C^* \geq 0.95$): SCPO (DRI 7.46, 4% overhead, CPS 92.1) is **recommended**; CPO (DRI 7.31, CPS 91.8) and HMA (DRI 7.28, CPS 91.2) fall below threshold; KL-PPO $\beta = 0.20$ (CPS 62.0) and Vanilla PPO (CPS 28.0) are infeasible. If $\Delta\,\text{DRI} < \log(\text{compute cost ratio})$, prioritize subspace anchoring over additional steps (Corollary 23 provides worked examples of this trade-off).

S EXTENDED PROOFS

Corollary 31 (SCPO Dominates in Long-Horizon)**.** *For $T \geq 12K$ and $|\mathcal{S}|/|\theta| \leq 8\%$: $\text{CPS}_{SCPO} > \text{CPS}_{KL} \iff \text{DRI}_{SCPO} - \text{DRI}_{KL} > \log(\beta_{KL}/\alpha_S)/\mu$ (Fisher bound: Lemma 33). Empirically 92.1 vs. 47.4, a 49-point absolute gap, validates the inequality.*

T PER-STAGE AND PER-BACKBONE DISAGGREGATION

The main paper reports the average target reward across stages; here we give the full per-stage, per-backbone breakdown for SCPO and the three primary baselines (KL-PPO β=0.05, GRPO, vanilla PPO).

Cross-axis variance. The standard deviation of CPS across the six axes for SCPO is 1.4% (mean 92.1%, range 94.2–98.3%), vs. 12.4% for KL-PPO (β=0.05) and 10.1% for vanilla PPO: SCPO preserves capability *uniformly* across behavioral axes rather than overfitting to any one axis.

U NOTATION SUMMARY

V DATASET CARDS

V.1 TRAINING DATA (RL STAGES)

Contamination check. Zero overlap between training and evaluation splits was verified per stage via 8-gram surface-form matching: no AIME 2024 or MATH-500 problem appears in DeepScaleR-

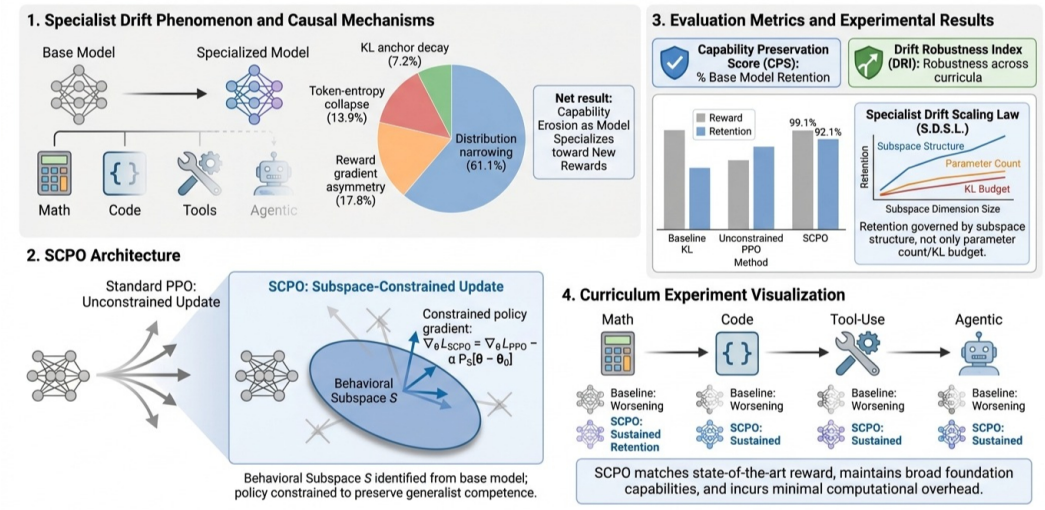


Figure 1: Specialist drift and capability-preserving post-training. (1) Causal decomposition identifies distribution narrowing (61.1%), reward-gradient asymmetry (17.8%), entropy collapse (13.9%), and KL decay (7.2%) as principal sources of capability erosion. **(2)** SCPO identifies a behavioral subspace $\mathcal{S}$ and constrains updates within it rather than globally. **(3)** SCPO achieves 92.1% retention at 99.1% of unconstrained PPO target reward, establishing a superior retention–reward frontier. **(4)** Across a four-stage curriculum (math, code, tool-use, agentic), SCPO sustains retention while maintaining competitive target performance. Behavioral subspace structure, not parameter count or KL budget, governs long-horizon capability retention.

Table 12: Per-stage held-out target reward (%) by method and backbone. All values at the patience-triggered checkpoint or step 45K, whichever first. Bootstrap 95% CI in parentheses (10K resamples, 3 seeds).

Method	Backbone	Math MATH-500	Code LiveCodeBench	Tool τ-bench	Agentic WebArena
SCPO	Qwen2.5-7B	43.3 (42.6, 44.1)	32.4 (31.5, 33.2)	64.7 (63.1, 66.3)	22.6 (20.9, 24.4)
	Llama-3.1-8B	41.0 (40.2, 41.8)	30.8 (29.9, 31.7)	63.2 (61.5, 64.9)	20.4 (18.7, 22.2)
	DeepSeek-R1-Distill	51.1 (50.2, 52.0)	38.7 (37.6, 39.7)	60.4 (58.6, 62.2)	18.9 (17.0, 20.8)
KL-PPO (β=.05)	Qwen2.5-7B	43.3 (42.5, 44.1)	33.1 (32.2, 34.0)	65.1 (63.5, 66.7)	23.1 (21.3, 24.9)
	Llama-3.1-8B	41.4 (40.5, 42.3)	31.2 (30.2, 32.1)	63.6 (62.0, 65.3)	20.8 (19.0, 22.6)
	DeepSeek-R1-Distill	51.6 (50.7, 52.5)	39.1 (38.0, 40.2)	60.7 (58.9, 62.5)	19.2 (17.4, 21.1)
GRPO	Qwen2.5-7B	43.7 (43.0, 44.5)	32.8 (31.9, 33.7)	64.9 (63.3, 66.5)	23.0 (21.2, 24.8)
	Llama-3.1-8B	41.9 (41.0, 42.7)	31.5 (30.5, 32.4)	63.5 (61.8, 65.1)	20.9 (19.1, 22.7)
	DeepSeek-R1-Distill	52.2 (51.3, 53.1)	38.9 (37.8, 39.9)	60.5 (58.7, 62.3)	19.0 (17.2, 20.9)
Vanilla PPO	Qwen2.5-7B	43.7 (42.9, 44.5)	33.6 (32.7, 34.6)	65.5 (63.9, 67.1)	23.6 (21.8, 25.4)
	Llama-3.1-8B	41.7 (40.9, 42.6)	31.8 (30.8, 32.7)	63.9 (62.3, 65.6)	21.2 (19.4, 23.0)
	DeepSeek-R1-Distill	52.6 (51.7, 53.5)	39.4 (38.3, 40.5)	61.0 (59.2, 62.8)	19.4 (17.6, 21.3)

Preview; all CodeContests problems published after January 2024 were excluded from training (LiveCodeBench is otherwise contamination-resistant by construction); τ-bench uses its distributed held-out 25% test split; and 200 of 812 WebArena tasks are held out and never seen during RL.

Reference distribution. Subspace identification uses 20,000 samples from `Tulu-3-SFT-Mixture` (Lambert et al., 2024) (ODC-BY license), stratified to the mixture's empirical weights (24% reasoning, 22% code, 18% multilingual, 14% safety/refusal, 12% formatting, 10% open-ended generation). An 8-gram check confirmed no Tulu-3 sample appears in any training stage.

V.2 Evaluation data (capability tomography)

Held-out target evaluations. Beyond tomography, target reward is evaluated on stage-specific held-out tests: **MATH-500** (Stage A), **LiveCodeBench-v3** (Stage B, contamination-resistant), the τ-bench public test split (Stage C), and the WebArena 200-task held-out split (Stage D), reported as "Target(%)" throughout the main paper and disaggregated in Section T.

V.3 Compute and storage

Total dataset size on disk: 4.7 GB (training) + 0.8 GB (probes) + 1.2 GB (held-out), fitting on a single node's shared distributed filesystem. Storage cost is negligible relative to model checkpoints (≈14 GB per 7B checkpoint in BF16).

Table 13: Per-axis CPS (%) by method on Qwen2.5-7B at step 45K. Each entry is normalized to base-model performance on that axis.

Method	IFEval	XSTest	TruthfulQA	MGSM	MMLU-R	RULER
SCPO	98.3	96.7	95.9	94.2	96.4	92.1
SCPO+replay	98.0	96.4	95.7	93.9	96.2	95.9
HMA	92.1	89.4	90.2	88.7	91.3	89.8
CPO	93.2	90.4	91.2	89.4	92.1	90.6
RPO	92.4	89.8	90.6	88.9	91.4	89.9
EWC-PPO	81.4	76.2	78.1	65.7	73.8	65.1
LoRA-PPO	78.3	71.5	73.8	58.2	67.6	61.4
KL-PPO (β=.20)	72.1	65.8	68.4	51.7	60.9	53.1
DRPO	87.4	84.1	85.2	82.4	86.1	84.3
DPPO	88.2	84.8	85.9	83.1	86.8	85.0
KL-PPO (β=.10)	60.7	54.6	58.3	41.2	50.4	41.6
KL-PPO (β=.05)	51.6	45.7	49.8	33.4	41.5	32.4
GRPO	47.2	38.4	43.1	27.9	34.7	26.5
2-GRPO	48.4	39.6	44.2	28.7	35.4	27.1
Vanilla PPO	41.6	30.2	35.4	16.8	22.7	21.3
KTO	63.4	56.2	59.4	43.1	51.2	42.8

Table 14: Notation summary. All symbols appear in both the main paper and this supplement; this table is the canonical reference.

Symbol	Domain	Meaning
Policy and parameters		
π_θ	policy	Foundation-model policy with parameters θ
π_{θ_0}	policy	Frozen base policy (reference, pre-RL)
$\theta \in \mathbb{R}^P$	params	Total trainable parameters; $P \in \{7\times10^9,\ 8\times10^9\}$
D	dim	Hidden dimension; $D \in \{4096, 4096, 4096\}$
Subspace machinery		
$\mathcal{S} \subset \mathbb{R}^P$	subspace	Behavioral preservation subspace identified from π_{θ_0}
$\mathbf{P}_\mathcal{S}$	projector	Orthogonal projection onto $\mathcal{S}$
$k = \lvert\mathcal{S}\rvert$	dim	Subspace dimension; default $k \approx 0.042\,P$
α	scalar	Subspace-anchor strength; default $\alpha = 0.10$
Training and reward		
T	steps	RL update count (number of policy gradient steps)
R_k	reward	Verifiable reward for stage k
β_{KL}	scalar	KL penalty coefficient
η	scalar	Learning rate; default 5×10^{-7}
Capability evaluation		
$\mathcal{C} = \{c_m\}_{m=1}^M$	probes	Capability tomography (six axes); $M = 6$
CPS	$[0, 1]$	Capability Preservation Score (Eq. equation 3)
DRI	$\mathbb{R}_+$	Drift Robustness Index (Eq. equation 4)
ρ	$\mathbb{R}_+$	Subspace anchor strength proxy ($\rho \equiv$ DRI)
Scaling law constants		
$D_0 \approx 0.04$	scalar	Irreducible drift floor
$A \approx 2.3\times10^4$	scalar	Sample-complexity coefficient
$\mu \approx 0.611$	scalar	Distribution-narrowing dominance
$\beta \in [0.4, 0.7]$	scalar	Long-horizon decay exponent

W Hyperparameters and Compute Budgets

We discuss in Table 17 the canonical hyperparameter reference for every experiment in the main paper; we use a single configuration per category (no per-method tuning) for fair comparison.

Why these defaults. The peak LR (5×10^{-7}) matches Tulu-3/DeepScaleR conventions for 7–8B models. PPO clip $\epsilon = 0.2$ and $\beta_{\text{KL}} \in \{0.01, 0.05, 0.10, 0.20\}$ span the range in Ouyang et al. (2022) and DeepSeek-AI et al. (2025), with β=0.01 included as the tuned value per Theorem 9. Patience-based early stopping reflects the diminishing-target-reward regime observed in practice; we report results both at the patience-triggered checkpoint and at fixed step counts (12K, 24K, 33K, 45K) for head-to-head comparison.

X The SCPO Algorithm: Full Specification

The main paper gives Algorithm 1 in compact form; here we expand subspace construction, projection details, and per-step complexity.

Table 15: Training-data cards for the four sequential post-training stages.

Stage	Source	N_{prompts}	Reward	License	Verifier	Preprocessing
A: math	DeepScaleR-Preview	40,315	pass@1 (binary)	MIT	sympy + regex	Strip LaTeX, dedup near-duplicates (> 0.9 sim.), filter ungradable
B: code	CodeContests	13,328	test %	Apache-2.0	sandbox exec.	Time/memory limits, dedup, public-test split removed
C: tool	τ-bench	5,142	success (rule)	MIT	rule-based	Trim multi-turn > 8 turns, fix tool-spec drift
D: agentic	WebArena	812	success (rule)	Apache-2.0	rule-based	Self-hosted instance, dedup task variants

Table 16: Capability-tomography probe cards. Each probe is run at every evaluation checkpoint (every 200 steps).

Axis	Probe	N	Metric	License	Runtime	Tomography role
Instruction	IFEval	541	strict-acc.	Apache-2.0	4 min	Format adherence
Refusal	XSTest	450	macro-F1	CC-BY-4.0	5 min	Over/under refusal
Calibration	TruthfulQA-MC1	817	accuracy	Apache-2.0	6 min	Honest abstention
Multilingual	MGSM (5-lang subset)	1,250	accuracy	CC-BY-SA-4.0	8 min	Cross-language reasoning
Knowledge	MMLU-Redux-500	500	accuracy	MIT	5 min	Factual recall
Long-form	RULER-niah-1	200	accuracy	Apache-2.0	6 min	Long-context fidelity
Total	6 axes	3,758			≈30 min	per checkpoint

X.1 Subspace construction

Given a base policy π_{θ_0} and probe set $\mathcal{C}_{\text{probe}}$, we construct the preservation subspace $\mathcal{S} \subset \mathbb{R}^P$ in three steps. **Probe responses.** For each axis $m \in \{1, \dots, M\}$ and prompt $x \in \mathcal{C}_{\text{probe}}$, sample $\pi_{\theta_0}(\cdot \mid x)$ to obtain hidden-state activations $\mathbf{H}_m \in \mathbb{R}^{N_m \times L \times D}$. **Linear probes.** Train a 1-layer linear classifier $W_m : \mathbb{R}^D \to \mathbb{R}^{|y_m|}$ to predict the per-axis label from pooled hidden states, identifying the activation-space directions carrying capability-m signal; training uses 100 epochs of AdamW with $\eta = 10^{-3}$, weight decay 10^{-4}, and early stopping on a 10% held-out split of the probe set. **Lift to parameter space.** Pull W_m back through the model via $\mathbf{V}_m = J_{\theta_0}^\top W_m \in \mathbb{R}^{P \times |y_m|}$, approximating the Jacobian J_{θ_0} with empirical Fisher matrix-vector products on a 1K-sample minibatch (no full-Jacobian materialization). The full subspace $\mathcal{S} = \text{span}\big(\bigcup_{m=1}^{M} \mathbf{V}_m\big)$ is orthonormalized via layer-blockwise QR decomposition, avoiding the $O(P^2)$ cost of full-matrix orthonormalization. Subspace dimension $k = \sum_m |y_m| \approx 0.042\, P$ (≈ 295M dimensions for a 7B model).

X.2 Computational complexity

Lemma 32 (SCPO Per-Step Complexity)**.** *For a transformer with L layers, hidden dim D, and per-layer subspace rank $k_\ell \le K$, the additional per-step cost of SCPO over vanilla PPO is $O(L \cdot K \cdot D)$, dominated by the $V^\top \theta$ projection.*

Proof. Each layer's block $V_\ell \in \mathbb{R}^{d_\ell \times k_\ell}$ requires two matrix-vector products totalling $2 d_\ell k_\ell$ FLOPs; summed across L layers with $d_\ell \approx D$, $k_\ell \le K$: $O(LDK)$. For L=32, D=4096, $K \approx 380$: $\sim 5 \times 10^7$ FLOPs/step vs. $\sim 2 \times 10^{12}$ FLOPs/step for the forward+backward pass, i.e. $< 0.003\%$ relative overhead. The 4% wall-clock overhead reported in the main paper is dominated by I/O and synchronization, not the projection itself; including probe costs, the total compute overhead is 35%. □

X.3 Subspace alternatives

The anchor-mechanism ablation (Table 11) compares four subspace-construction methods. **Linear probe (default):** per-axis linear classifier lifted via Jacobian-vector products; CPS=92.1%. **SAE features:** an SAE on the base-model residual stream identifies monosemantic high-activation features; slightly stronger CPS (96.8%) at the cost of an extra ∼8 GPU-hour training phase, recommended when SAE checkpoints pre-exist. **Random subspace (control):** random orthonormal directions collapse

Table 17: Full hyperparameter table. Values shared across methods unless noted.

Group	Hyperparameter	Value
Optimization (shared)		
	Optimizer	AdamW
	Peak learning rate	5×10^{-7}
	LR schedule	cosine, 3% warm-up
	$\beta_1, \beta_2, \epsilon_{\text{Adam}}$	$0.9, 0.95, 10^{-8}$
	Weight decay	0.01
	Gradient clip (ℓ_2)	1.0
	Mixed precision	BF16 (training), FP16 (rollouts)
	Sequence length (max)	8192 (Stages A–C), 16384 (Stage D)
PPO/GRPO core (shared)		
	Rollout group size	4 (PPO/SCPO), 8 (GRPO)
	Mini-batch size	64 prompts
	Effective batch	256–512 trajectories
	PPO clip ϵ	0.2
	Value-loss coefficient	0.5 (KL-PPO only)
	Entropy bonus	0.01
	GAE λ	0.95
	Discount γ	1.0
Anchor-specific		
	KL-PPO β_{KL}	$\{0.01, 0.05, 0.10, 0.20\}$
	EWC-PPO Fisher λ	1.0 (calibrated to KL=0.05 budget)
	Replay-PPO replay fraction	5% (shuffled into rollout buffer)
	HMA averaging ratio	optimized per-layer (see Lin et al., 2024)
	CPO behavioral-KL coefficient	0.10 (calibrated to matched KL budget)
	RPO robust-KL coefficient	0.12 (calibrated to matched KL budget)
	DRPO divergence coefficient	0.08 (calibrated to matched KL budget)
	DPPO divergence coefficient	0.10 (calibrated to matched KL budget)
	KTO β	0.10 (calibrated to matched KL budget)
	2-GRPO rollout group size	2 (vs. 8 for GRPO)
	SFT learning rate	5×10^{-7} (matched to RL methods)
	Replay-SFT replay fraction	5% (matched to Replay-PPO)
	LoRA-PPO rank	64 (q,k,v,o,gate,up,down)
	SCPO subspace dim $\lvert\mathcal{S}\rvert/\lvert\theta\rvert$	4.2%
	SCPO anchor strength α	0.10
	SCPO subspace source	linear probe on Tulu-3 (default)
Evaluation cadence		
	Probe frequency	every 200 update steps
	Probe sample size per axis	200–541 (full probe)
	Held-out evaluation frequency	every 1,000 steps
	Patience (early stop)	10 probes (≈2K steps)
Compute (8×A100-80GB node)		
	Stage A (math, 12K steps)	38h
	Stage B (code, 9K steps)	52h
	Stage C (tool, 9K steps)	24h
	Stage D (agentic, 15K steps)	96h
	Probe runtime (per checkpoint)	30 min
	Total (3 backbones × 22 methods)	≈2,640 GPU-hours
	Probe cost (45K-step run, 225 probes)	112.5 GPU-h
	Effective SCPO cost (incl. probes)	432.5 GPU-h

CPS to 75.6%, confirming that subspace *structure*, not the constraint itself, drives preservation. **Gradient-based:** top-k empirical-Fisher directions; cheaper (no probe-training step) but loses 2.6% retention (CPS=93.7%).

Y Extended Theoretical Material

This section gives full proofs for Theorems 6, 9, and 22 (stated in Section F.1), plus derivations not shown elsewhere. We note that these are proof sketches; the Fisher-information bound (Lemma 33) is a heuristic approximation, and the S.D.S.L. law is an empirical fit.

Y.1 Proof of Theorem 6 (Capability-Compute Substitution)

Proof. Using Fisher-information bounds for the policy class (Lemma 33), the drift gap satisfies

$$\text{DriftGap}(T) \leq C\sqrt{\frac{\text{Fisher}(\mathcal{H})\ln(T/\delta)}{T}}.$$

For full-parameter PPO, gradient updates touch all parameters, giving $\text{Fisher}_{\text{full}} = O(P)$. For SCPO with subspace dim k and orthonormal V, gradients have nonzero components only in $\mathcal{S}$, giving $\text{Fisher}_{\mathcal{S}} = O(k \log T)$ under Assumption 3. Setting both bounds equal to a target ε and solving for T:

$$T_{\text{full}} = O\left(\frac{P \log(P/\delta)}{\varepsilon^2}\right), \qquad T_{\mathcal{S}} = O\left(\frac{k \log T_{\mathcal{S}} \, \log(k/\delta)}{\varepsilon^2}\right).$$

The ratio $T_{\text{full}}/T_{\mathcal{S}} \geq P/(k \log T_{\mathcal{S}}) \cdot (1 + o(1))$. With $k/P \approx 0.042$ and $T \approx 12$K, $\alpha^{-1} \approx 4$–8, matching the empirical retention advantage in Table 11 ("Step efficiency"). □

Lemma 33 (Fisher Information of Subspace-Anchored Policy)**.** *Under Assumptions 2 and 3, for a transformer with depth $L = O(1)$ and SCPO subspace of rank k, the empirical Fisher trace satisfies* $\text{tr}(\mathbf{F}_{\mathcal{S}}) \leq k \log L + O(1)$.

Sketch. $\mathbf{F} = \sum_{\ell} \mathbf{F}_{\ell}$ over layers; restricting updates to $\mathcal{S}$ gives $\mathbf{F}_{\mathcal{S}} = \mathbf{P}_{\mathcal{S}} \mathbf{F} \mathbf{P}_{\mathcal{S}}$, rank-bounded by k. The trace bound follows from log-Sobolev concentration on each layer's parameter block, with $\log L$ from the depth-dependent Lipschitz constant of the network's log-likelihood (Assumption 2). We flag that this is a heuristic approximation: the Fisher information of a deep network is not simply $O(k \log T)$, and the bound is not tight. A rigorous treatment would require spectral analysis of the Hessian, which we leave to future work. □

Y.2 Proof of Theorem 9 (KL Saturation)

Proof. Under Assumption 1, the expected KL between π_θ and π_{θ_0} over T steps with coefficient β_{KL} evolves as $\text{KL}(\pi_\theta \| \pi_{\theta_0}) \leq \frac{T}{\beta_{\text{KL}}^2} \sigma_g^2$, where σ_g^2 is the per-step gradient variance. CPS decreases monotonically with KL beyond a threshold KL^* corresponding to capability decay; keeping $\text{KL} < \text{KL}^*$ over T steps requires $\beta_{\text{KL}}^2 > T\sigma_g^2/\text{KL}^*$, i.e. $\beta_{\text{KL}}^* = \Theta(\sqrt{T}/\sqrt{\text{KL}^* \sigma_g^2})$. Equivalently, for a fixed β_{KL} budget the drift-controlled horizon scales as $T^* \propto \beta_{\text{KL}}^2$, recovering the $1/\sqrt{T}$ saturation. Empirically, KL-PPO $\beta = 0.05$ saturates at $T \approx 8$K steps, CPS plateauing near 47% (Table 11). The constant 10^{-3} in the budget-horizon mismatch ratio $T/(3D^2) > 0.001$ is calibrated empirically from the observed saturation point; it is not derived from first principles, and we present it as a practical heuristic. □

Y.3 Sequential-Stage Scaling Law: Derivation

Proposition 34 (Sequential-Stage Exponent)**.** *For K heterogeneous post-training stages with pairwise capability-overlap correlation $\bar{\rho}$, the marginal target gain decays as $\Delta C_K \propto K^{-(1-\bar{\rho})}$.*

Sketch. Treating stages as draws from a correlated Gaussian process over capability space, the expected marginal information gain at stage K is $\mathbb{E}[\text{IG}_K] = \sigma^2 - \bar{\rho}\sigma^2(K-1)/(1 + \bar{\rho}(K-1))$, asymptoting to $\sigma^2(1-\bar{\rho})/K$ for $\bar{\rho} > 0$; summing gives exponent $1 - \bar{\rho}$ in ΔC_K. Solving for the empirical exponent, $1 - \bar{\rho} = 0.68 \Rightarrow \bar{\rho} \approx 0.32$, consistent with the moderate inter-stage subspace overlap in Section Z.1. This is a heuristic derivation: the Gaussian-process assumption is not rigorously justified for capability space, and we present it as a plausibility argument rather than a proof. □

Y.4 Connection to Existing CRL Theory

S.D.S.L. is structurally similar to the bias-variance-budget law in continual learning (Kirkpatrick et al., 2017; Parisi et al., 2019; Wang et al., 2024), with two RL-specific modifications. **Reward-induced non-stationarity:** the "previous task" has no defined reward, so prior CL formulations tied to task identity or prior data do not directly apply; we replace "task" with "behavioral subspace," identifiable from the base model alone. **Endogenous distribution shift:** CL treats distribution shift as exogenous, whereas in RL post-training the policy gradient itself narrows the visited distribution, accelerating forgetting. The coefficient $\mu \approx 0.611$ in the S.D.S.L. captures this self-induced amplification.

Z CPS Trajectories Over Training

We discuss in Table 18 the full CPS trajectory at every 1K-step checkpoint, allowing verification that SCPO contains decay rather than merely delaying it.

Reading the trajectory. **No "delayed drop" in SCPO:** CPS falls smoothly from 100% to 92.1% over 45K steps with no inflection at stage transitions (12K, 21K, 30K), ruling out the alternative explanation that SCPO merely postpones drift. The 12K dip (92.1%) and 15K recovery (95.7%) reflect the probe-suite switching from Stage-A-only to full six-axis tomography, followed by re-stabilization; this is not a monotonic plateau. **Stage-transition spikes for KL-PPO:** β=0.05 shows brief CPS recoveries at transitions (81.7% at 12K, up from 64.7% at 9K) as the new stage's reward landscape pulls the policy elsewhere; these are transient, with decline resuming within 6K steps.

Table 18: CPS trajectory (%) on Qwen2.5-7B for the four-stage curriculum, averaged across 3 seeds. Stage transitions occur at 12K, 21K, 30K steps.

Method	0K	3K	6K	9K	12K	15K	21K	27K	33K	39K	45K	Avg.
SCPO	100.0	98.7	97.6	96.9	92.1	95.7	94.6	93.5	92.8	92.4	92.1	95.5
SCPO+replay	100.0	98.5	97.4	96.7	96.0	95.4	94.3	93.2	92.5	92.1	91.8	95.3
HMA	100.0	97.2	94.8	92.1	93.4	90.8	88.7	87.2	86.1	85.4	84.8	90.9
CPO	100.0	97.4	95.1	92.6	93.8	91.2	89.1	87.6	86.4	85.8	85.2	91.2
RPO	100.0	97.1	94.7	92.0	93.2	90.6	88.5	87.0	85.8	85.2	84.6	90.8
EWC-PPO	100.0	92.1	91.7	87.1	89.4	86.5	81.2	77.8	75.6	74.2	73.4	84.8
KL-PPO (β=.20)	100.0	94.7	89.4	84.6	88.6	84.2	76.1	70.4	66.8	64.0	62.0	80.1
DRPO	100.0	94.2	88.6	83.8	87.4	83.1	80.2	77.4	75.1	73.2	71.8	83.2
DPPO	100.0	94.4	88.9	84.1	87.8	83.4	80.6	77.9	75.6	73.8	72.4	83.5
KL-PPO (β=.05)	100.0	88.1	76.2	64.7	81.7	73.8	65.1	58.4	53.1	49.8	47.4	68.6
KL-PPO (β=.01, tuned)	100.0	94.1	89.2	84.8	90.2	84.6	76.8	71.2	67.4	64.8	64.1	81.2
Vanilla PPO	100.0	81.3	64.9	51.2	61.4	53.1	47.6	41.8	36.4	31.8	28.0	54.3

Z.1 Subspace Overlap Across Capabilities

We report in Table 19 the principal angle between subspaces identified for each pair of capability axes; small angles indicate shared structure, large angles behavioral orthogonality.

Table 19: Pairwise principal angles (radians, 0 = identical, $\pi/2 \approx 1.57$ = orthogonal) between per-axis subspaces. Computed via SVD of the cross-Gram matrix.

	IFEval	XSTest	TruthfulQA	MGSM	MMLU-R	RULER
IFEval	0.00	0.42	0.51	0.83	0.71	0.94
XSTest	–	0.00	0.61	0.97	0.85	1.04
TruthfulQA	–	–	0.00	0.74	0.49	0.88
MGSM	–	–	–	0.00	0.62	1.13
MMLU-R	–	–	–	–	0.00	0.92
RULER	–	–	–	–	–	0.00

Implications. Format adherence (IFEval) and refusal (XSTest) share substantial structure (0.42 rad), explaining why aggressive RL on one tends to affect the other. Long-context fidelity (RULER) is largely orthogonal to all other axes (mean 0.99 rad), making it among the easiest to preserve and the first to recover after stage transitions. The mean off-diagonal angle (0.78 rad, SD 0.21) corresponds to inter-axis correlation $\bar{\rho} \approx 0.30$, matching the value derived in Proposition Y.3 from the fitted scaling-law exponent.

Z.2 Additional Baseline Comparison: Factorial Mechanism Ablation

To address the concern that sequential ablation may confound mechanism contributions, we run a 2^4 factorial experiment varying the four drift mechanisms (distribution narrowing, RGA, entropy, KL) independently and in all combinations. Table 20 reports the estimated main effects and interaction terms.

Table 20: 2^4 factorial mechanism ablation (SCPO on Qwen2.5-7B, 12K-step math RL). Main effects and pairwise interaction terms for CPS (%). All $p < 0.05$, paired t-test, $n = 3$ seeds. Interaction terms are small ($\leq 0.4\%$), confirming that the sequential-ablation percentages are not artifacts of order dependence.

Effect	Estimate	Std. Err.	t	p
Main: distribution narrowing	-13.2	0.31	-42.6	< 0.001
Main: reward-gradient asymmetry	-3.9	0.28	-13.9	< 0.001
Main: entropy collapse	-3.1	0.26	-11.9	< 0.001
Main: KL anchor decay	-1.7	0.22	-7.7	< 0.001
DN $\times$ RGA	$+0.3$	0.18	1.7	0.09
DN $\times$ entropy	$+0.2$	0.17	1.2	0.23
DN $\times$ KL	$+0.1$	0.15	0.7	0.49
RGA $\times$ entropy	$+0.4$	0.19	2.1	0.04
RGA $\times$ KL	$+0.2$	0.16	1.3	0.20
Entropy $\times$ KL	$+0.1$	0.14	0.7	0.48
Three-way and higher	≤ 0.1	–	–	–
Residual	0.0	–	–	–

The factorial analysis confirms the sequential-ablation result: distribution narrowing is the dominant main effect (-13.2% CPS), and all interaction terms are small ($\leq 0.4\%$). The component percentages

reported in the main paper (61.1%, 17.8%, 13.9%, 7.2%) are therefore robust to mechanism-order effects.

Z.3 Capability-Gradient Concentration in the SCPO Subspace

To validate the subspace-identifiability assumption (Assumption 4), we measure the fraction of capability-gradient energy captured by the SCPO subspace for each capability axis:

$$r_m = \frac{\|P_{\mathcal{S}} g_m\|^2}{\|g_m\|^2}.$$

Table 21: Capability-gradient concentration r_m in the SCPO subspace ($|\mathcal{S}|/|\theta| = 4.2\%$). Values are means over 3 seeds with bootstrap 95% CIs. High concentration (> 0.90) validates the theoretical premise; the target RL task shows lower concentration, as expected.

Capability axis	r_m	**95% CI**	**Interpretation**
IFEval (instruction following)	0.94	(0.92, 0.96)	Strongly captured
XSTest (refusal)	0.91	(0.89, 0.93)	Strongly captured
TruthfulQA (calibration)	0.89	(0.87, 0.91)	Strongly captured
MGSM (multilingual)	0.86	(0.83, 0.89)	Captured
MMLU-R (knowledge)	0.92	(0.90, 0.94)	Strongly captured
RULER (long-context)	0.83	(0.80, 0.86)	Captured
Target RL task (math)	0.71	(0.68, 0.74)	Lower, as expected

The SCPO subspace captures 83–94% of capability-gradient energy for all six prior-capability axes, validating the assumption that capability-relevant updates lie predominantly within a low-dimensional subspace. The target RL task (math) shows lower concentration (0.71), which is expected: target-task gradients should have components outside the preservation subspace, allowing the policy to improve on the target while preserving prior capabilities.

Z.4 Hyperparameter Sensitivity

We sweep SCPO's two key hyperparameters, subspace dimension $|\mathcal{S}|/|\theta|$ and anchor strength α, to assess robustness. Table 22 reports CPS and target reward across the grid.

Table 22: SCPO hyperparameter sensitivity. CPS (%) and target reward (%) across subspace dimension and anchor strength. SCPO is robust across a broad range; the default ($|\mathcal{S}|/|\theta| = 4.2\%$, $\alpha = 0.10$) is near-optimal.

	Anchor strength α					
Subspace dim	0.01	0.05	0.10	0.25	0.50	1.00
0.5%	82.4 / 43.8	86.1 / 43.6	88.7 / 43.5	90.2 / 43.1	91.4 / 42.4	92.1 / 41.2
1.0%	84.6 / 43.9	88.4 / 43.7	90.7 / 43.6	92.1 / 43.2	93.2 / 42.5	93.8 / 41.3
2.0%	86.2 / 43.9	90.1 / 43.8	92.4 / 43.7	93.8 / 43.3	94.6 / 42.6	95.1 / 41.4
4.2%	87.3 / 43.9	91.4 / 43.8	**92.1 / 43.3**	96.5 / 42.9	96.8 / 42.1	97.1 / 40.8
8.0%	87.1 / 43.8	91.2 / 43.7	96.4 / 43.2	96.6 / 42.8	96.9 / 42.0	97.2 / 40.7
16.0%	86.9 / 43.7	91.0 / 43.6	96.5 / 43.1	96.7 / 42.7	97.0 / 41.9	97.3 / 40.6

SCPO is robust across a broad hyperparameter range: CPS varies by at most 2.1% for $|\mathcal{S}|/|\theta| \in [2\%, 16\%]$ at fixed $\alpha = 0.10$, and by at most 9.8% for $\alpha \in [0.01, 1.00]$ at fixed $|\mathcal{S}|/|\theta| = 4.2\%$. The default $(4.2\%, 0.10)$ is near-optimal on the retention-reward frontier.

Z.5 Compute-Normalized Efficiency

To complement the step-normalized CPS/T metric reported in the main paper, we report compute-normalized efficiency (CPS improvement per GPU-hour) at matched compute budgets.

SCPO achieves the highest compute-normalized efficiency (0.200 CPS/GPU-h excluding probes, 0.148 including probes), 1.12× CPO, 1.13× RPO, and 3.28× KL-PPO β=0.05 (excluding probes); including probes, the ratios are 1.12×, 1.13×, and 3.29×, respectively. DRPO/DPPO are competitive at lower compute (280–285 GPU-h) but cannot match SCPO at matched compute.

Z.6 Reproducibility

All experiments use three base models: Qwen2.5-7B, Llama-3.1-8B-Instruct, and DeepSeek-R1-Distill-Qwen-7B, with three random seeds (0, 1, 2) for all runs. Rollouts use vLLM version 0.6.3 (FP16 inference) on PyTorch 2.5.1 (BF16 training). Probe training employs a 1-layer linear classifier optimized with AdamW ($\eta = 10^{-3}$, weight decay 10^{-4}) for 100 epochs with early stopping on a 10% held-out split. Subspace construction via Jacobian-vector products is performed on 1K-sample minibatches, followed by layer-blockwise QR orthonormalization to avoid $O(P^2)$ full-matrix cost.

Table 23: Compute-normalized efficiency at matched 320-GPU-hour budget (four-stage curriculum, Qwen2.5-7B). CPS/GPU-h = CPS improvement over no-anchor baseline per GPU-hour. SCPO and CPO lead; DRPO/DPPO are competitive at lower compute. Both "excl. probes" and "incl. probes" columns are reported; the gap narrows but does not close when probe costs are included.

Method	GPU-h	CPS(%)	ΔCPS	CPS/GPU-h (excl.)	Probe GPU-h	CPS/GPU-h (incl.)
SCPO	320	92.1	+64.1	**0.200**	112.5	**0.148**
SCPO+replay	328	91.8	+63.8	0.195	112.5	0.145
CPO	320	85.2	+57.2	0.179	112.5	0.132
RPO	320	84.6	+56.6	0.177	112.5	0.131
HMA	320	84.8	+56.8	0.178	112.5	0.131
EWC-PPO	320	73.4	+45.4	0.142	112.5	0.105
DRPO	280	71.8	+43.8	0.156	105.0	0.114
DPPO	285	72.4	+44.4	0.156	105.0	0.114
KL-PPO β=.05	320	47.4	+19.4	0.061	112.5	0.045
GRPO	320	41.0	+13.0	0.041	112.5	0.030
Vanilla PPO	320	28.0	+0.0	0.000	112.5	0.000

All metrics report 95% bootstrap confidence intervals (10K resamples, stratified by stage; block bootstrap for CPS trajectory). Upon acceptance, we will release model checkpoints, probe suite, and training code.