RL FINE-TUNING HEALS OOD FORGETTING IN SFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) is a standard post-training recipe for improving Large Language Models (LLM) reasoning, but why it works remains unclear. We revisit the common claim that "SFT memorizes, RL generalizes" through checkpoint-wise analyses of in-distribution (ID) and out-of-distribution (OOD) reasoning. We find that OOD performance often peaks early during SFT and then declines despite continued improvement in ID reasoning. RL typically does not surpass this early SFT peak; rather, it restores OOD capability lost during later SFT, and only from a bounded range of SFT checkpoints. Further spectral analysis and direction interventions show that forgetting and recovery are correlated with singular-vector directions, whereas singular-value substitutions have much smaller effects. These findings suggest a more precise view of post-training dynamics: SFT can forget, RL can recover, and controlling singular-vector rotation may improve OOD robustness.

1 INTRODUCTION

Supervised Fine-Tuning (SFT) is the dominant method for adapting Large Language Models (LLMs) after pretraining (Howard & Ruder, 2018; Radford et al., 2018). Reinforcement Learning Fine-Tuning (RLFT), particularly when initialized from an SFT checkpoint, has become a standard second stage for mathematical reasoning (DeepSeek-AI, 2025; Hugging Face, 2025). Related adaptation work spans code generation (Jiang et al., 2024), embodied reasoning (Lin et al., 2025; Li et al., 2025), and video reasoning (Shi et al., 2025).

Prior work establishes several complementary pieces of this pipeline. Endpoint comparisons show that SFT supplies task and format competence whereas outcome-based RL can transfer better to unseen variants (Chu et al., 2025). Large-scale handoff studies show that high SFT scores can misrank initializers for downstream RL and propose held-out generalization loss or pass@large-k as stronger predictors (Kang et al., 2026). SFT-side methods improve generalization by changing token weighting, reference anchoring, or the training distribution (Wu et al., 2026; Zhu et al., 2026; Zhao et al., 2026; Zhang et al., 2026). Base-relative analyses further show that online RL favors solutions closer to the starting policy in KL and therefore preserves prior capabilities better than SFT (Shenfeld et al., 2026), while fixed-initializer RLVR studies ask whether RL improves access to reasoning paths already supported by a policy (Yue et al., 2025). These results explain important endpoint, selection, and retention effects while leaving open the trajectory-level question studied here: *how does the OOD effect of downstream RL evolve as its initializer moves through an SFT trajectory?*

Our question and distinction. Let θ_k denote SFT checkpoint k and let $\mathcal{R}$ denote the downstream RL operator. Our central object is the paired checkpoint-wise effect

$$\Delta_{\mathcal{R}}^{\mathrm{OOD}}(k) = \mathrm{Acc}_{\mathrm{OOD}}(\mathcal{R}(\theta_k)) - \mathrm{Acc}_{\mathrm{OOD}}(\theta_k).$$

This differs from comparing terminal SFT with terminal RL, ranking independently trained SFT candidates by their final post-RL scores, or analyzing RL from one fixed initializer. Measuring $\Delta_{\mathcal{R}}^{\mathrm{OOD}}(k)$ across a dense SFT trajectory exposes the complete handoff geometry: an early region where RL hurts, an intermediate region where RL heals forgotten OOD capability, and a late region where recovery disappears again. Our contribution is the discovery of both an *entry boundary* and an *exit boundary*, together with the early OOD peak that the intermediate RL stage largely restores, extending the observation that late SFT can be a poor initializer.

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.

We perform full-parameter SFT and RLFT and study in-distribution (ID) and out-of-distribution (OOD) reasoning across LLaMA-3.2-11B-Vision (Grattafiori et al., 2024) and Qwen-2.5 models (Team, 2024). We track dense checkpoint trajectories on controlled arithmetic, navigation, and cross-concept linear-algebra shifts, and pair them with Qwen/DAPO evaluations on public benchmarks. The controlled settings isolate explicit rule changes and make dense sweeps feasible; the public evaluations extend the analysis across cross-task shifts.

Contributions. **SFT forgets.** OOD reasoning peaks early during SFT and then declines while ID loss and accuracy continue to improve, revealing an ID–OOD divergence that terminal checkpoints and standard ID metrics conceal. **RL recovers.** The signed OOD effect of RL changes from negative to positive and back to negative along SFT. Three-seed LLaMA/PPO trajectories and a dense Qwen/DAPO sweep reproduce this bounded handoff, and the early/late operating-point comparison shows that late-SFT+RL nearly regains the early OOD peak while retaining substantially stronger ID capability. **Rotation matters.** Singular-value substitutions have little functional effect, whereas singular-vector substitutions produce large ID and OOD shifts. Base-relative and rank-matched interventions further show that recovery is direction-sensitive and consistently contains a component toward pretrained parameter and functional subspaces. A protected-SFT experiment provides a direct design implication: suppressing rotation in top-k subspaces preserves more late-SFT OOD capability.

2 PRELIMINARIES

Supervised Fine-Tuning (SFT). SFT adapts a pre-trained model π_θ to a specific task using a labeled dataset $\mathcal{D} = \{(x, y)\}$ (Howard & Ruder, 2018; Radford et al., 2018). The standard objective is to minimize the negative log-likelihood (NLL) of the target outputs given the inputs:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{(x,y)\in\mathcal{D}} \log \pi_\theta(y \mid x) \tag{1}$$

Reinforcement Learning (RL) Fine-Tuning In contrast to SFT, RL essentially fine-tunes the model by optimizing the policy π_θ based on a reward signal $R(\cdot)$. The general objective is to maximize the expected reward of the response generated by the model, $\max_\theta \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[R(x, y)]$. The reward function $R(x, y)$ encodes desired behavior through human preferences (Ouyang et al., 2022), model-based quality judgments (Wang et al., 2023), or constitutional principles (Bai et al., 2022). In this paper, we employ Proximal Policy Optimization (PPO) (Schulman et al., 2017), a popular RL algorithm that stabilizes training by optimizing a clipped surrogate objective. The PPO objective is:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min\left(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right], \tag{2}$$

where $r_t(\theta) = \frac{\pi_\theta(y_t|x,y_{<t})}{\pi_{\theta_{\text{old}}}(y_t|x,y_{<t})}$ is the probability ratio for state $(x, y_{<t})$ and action y_t at step t, A_t is the advantage estimate, and ϵ is a hyperparameter that constrains the policy update step to avoid excessive shift of policy. The advantage A_t measures how much better (or worse) taking action y_t in state $(x, y_{<t})$ is compared to the average action at that state, as estimated by a value function $V_\phi(x, y_{<t})$. Intuitively, A_t is positive when an action yields higher return than expected and negative otherwise.

3 EVALUATION AND ANALYSIS

Our checkpoint-wise evaluation tracks the evolution of LLM OOD reasoning through the SFT and RL stages. More specifically, in Section 3.1, we introduce the experimental settings, including the tasks, models and evaluation methods. In Section 3.2, we present the results and conduct detailed analysis on ID and OOD reasoning performance.

3.1 EVALUATION SETTINGS

Task Description We use *GeneralPoints, Navigation* (Chu et al., 2025) and *Rank-Determinant Computation* (Sun et al., 2025) to evaluate the arithmetic, spatial and cross-concept math reasoning abilities of models. Detailed task definitions and prompts appear in Appendix E.1; the main paper uses *GeneralPoints* as its running example. The *GeneralPoints* environment (Chu et al., 2025) is

instantiated on top of the *Points24* environment (Zhai et al., 2024). Each state s contains four poker cards, described in text directly. The goal is to produce an equation that equals a target number (24 by default), with four numbers from the cards used only once. Particularly, the cards $'J, Q, K'$ are all interpreted as the same number 10 in the original setting (for training). For example, provided with four cards $[5, 4, 10, 7]$, we aim to output the equation *(7-5)*10+4* as the desired output.

Evaluation of OOD Generalization We modify the rule of *GeneralPoints* as in Chu et al. (2025) and test both ID and OOD reasoning performance. Training interprets $'J, Q, K'$ as 10, while the shifted rule maps them to 11, 12, and 13. Holding the task format fixed makes this a controlled test of arithmetic transfer under a rule change. The public math/reasoning evaluations in Section 3.3 extend the analysis to cross-task shifts. We record checkpoint-wise performance to expose how ID and OOD accuracy evolve; Appendix E.1 details the corresponding shifts for Navigation and Rank-Determinant Computation.

Why Controlled and Public Evaluations Are Complementary. Controlled tasks provide diagnostic environments with explicit ID/OOD rule shifts and dense feasible sweeps, separating format alignment from reasoning-rule transfer while reducing pre-training-contamination ambiguity. We pair them with a public math/reasoning pipeline using Open-R1 SFT (Hugging Face, 2025) followed by GRPO or DAPO-math RL (Shao et al., 2024; Yu et al., 2025). Together, the controlled and public settings test the SFT-forgetting/RL-recovery pattern across explicit rule changes and cross-task shifts; Section 3.3 reports the public-benchmark evidence.

Models and Setup We use LLaMA-3.2-11B-Vision (Grattafiori et al., 2024) and Qwen-2.5 models (Team, 2024). Following the standard two-stage post-training pipeline (DeepSeek-AI, 2025), we first train with SFT and then initialize RL from selected SFT checkpoints. For *GeneralPoints*, we adopt the task and prompt format from Chu et al. (2025). We denote the terminal checkpoints as SFT_{End} and RL_{End}. The dense LLaMA/PPO sweep initializes RL at {90, 140, 200, 300, 400, ..., 1600}; the public Qwen2.5-3B/DAPO sweep evaluates 20 normalized SFT steps from 100 to 2000. Behavioral trajectory means and SDs use three independent training seeds and a binary verifier reward (1 correct, 0 incorrect), with an unchanged decoder within each comparison. Evaluation-example and layer-block confidence intervals, together with five-seed random-direction controls, are labeled separately. Detailed computational resources and training settings appear in Appendix E.2.

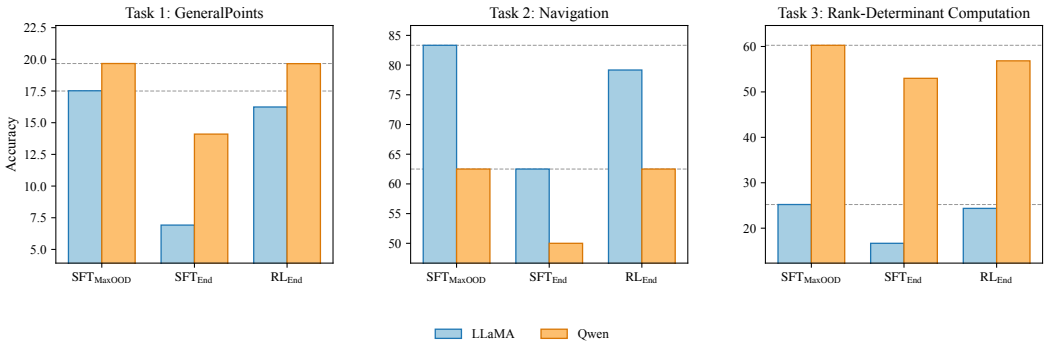
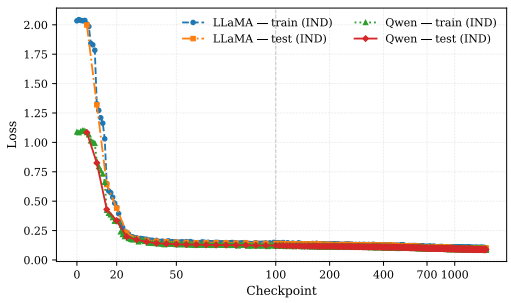


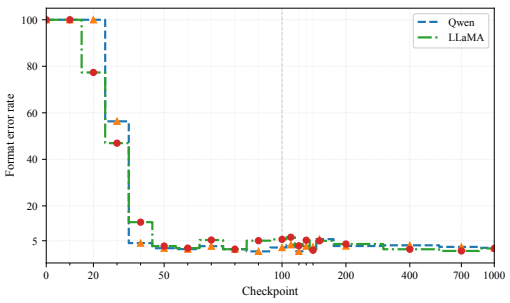
Figure 1: Comparison of OOD performance in three tasks for LLaMA and Qwen at different checkpoints ($\text{SFT}_{\text{MaxOOD}}$, SFT_{End} and RL_{End}).

3.2 Results and Analysis

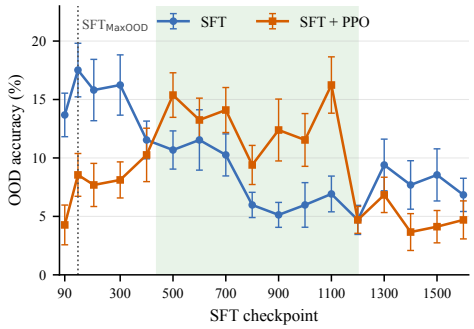
From endpoint contrast to checkpoint-wise recovery. Prior endpoint comparisons show that SFT stabilizes output before RL and that RL can improve the OOD accuracy of a terminal SFT model (Chu et al., 2025). Figure 1 reproduces this endpoint result across three controlled tasks. Our checkpoint-wise analysis measures how downstream RL changes OOD accuracy relative to each SFT initializer. This paired effect is exactly $\Delta_{\mathcal{R}}^{\text{OOD}}(k)$ from the Introduction. Unlike final-score comparisons across independently trained SFT candidates (Kang et al., 2026), the dense sweeps vary the handoff checkpoint within each experimental setup.



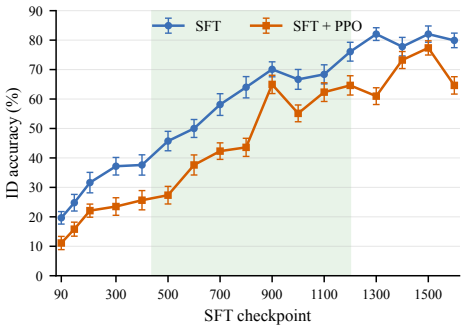
(a) In-distribution training and test loss



(b) Format error



(c) Evolution of OOD test accuracy.



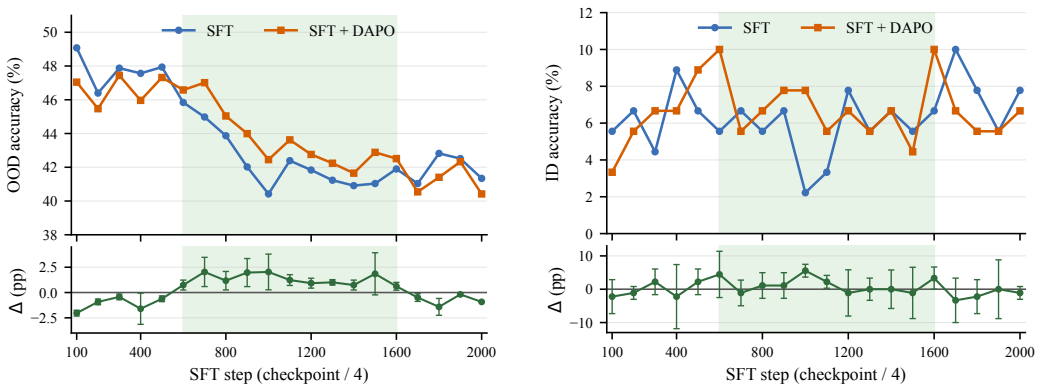
(d) Evolution of ID test accuracy.

Figure 2: (2a) Training and ID test loss, and (2b) format error during SFT. (2c) OOD and (2d) ID accuracy for LLaMA-3.2-11B-Vision/PPO/GeneralPoints across SFT initializations. Accuracy points are means across three independent training runs; error bars show one standard deviation. The shaded interval marks the estimated RL-recovery region.

Tracking the full SFT trajectory reveals two structures that endpoint evaluation misses. First, an earlier SFT checkpoint can outperform both terminal SFT and terminal RL on OOD accuracy; we denote the best observed trajectory checkpoint by $\text{SFT}_{\text{MaxOOD}}$ and use it as an oracle trajectory reference. ID training/test loss alone fails to identify this checkpoint in Figure 2a, motivating direct checkpoint-wise OOD analysis and the validation-based selector in Appendix C.2. Second, the sign of the paired RL effect changes twice along SFT. Qwen/GRPO improves OOD accuracy at checkpoints 400–1100 and lowers it at 100–300 and 1200–1600; DAPO improves all four reported OOD benchmarks at checkpoints 600 and 1400 and yields mixed or negative changes at the outer checkpoints (Appendix Figure 10 and Tables 8–9). The dense LLaMA/PPO and Qwen/DAPO trajectories below establish the same entry–recovery–exit geometry under independent model, data, and RL settings.

SFT forgets. Figure 2a shows the SFT training and ID test losses, Figure 2b shows format error, and Figures 2c–2d show the three-seed LLaMA accuracy trajectories. Format error falls to low single digits by checkpoint 50, while OOD verifier accuracy continues from 12.39% at checkpoint 50 to 17.52% at SFT_{140}. Later SFT increasingly specializes to the ID rule. After $\text{SFT}_{\text{MaxOOD}}$, OOD accuracy falls as ID loss decreases and ID accuracy rises. We call this divergence **OOD forgetting**: continued ID improvement conceals a systematic loss of OOD capability. Complementary loss curves show OOD loss rising while ID loss keeps falling, and the early verifier-error decomposition shows illegal-format errors reaching zero by checkpoint 70 as OOD accuracy grows from 0 to about 17% (Appendix Figure 12 and Appendix G.6). Across three seeds, the LLaMA forgetting/recovery magnitudes are $10.60\pm2.99/9.32\pm3.38$ pp, and the Qwen/GRPO mean differences are $5.57/5.56$ pp; resampling places 98.8% of LLaMA peaks in checkpoints 90–300 and the central 90% Qwen range in 110–400 (Appendix Tables 3 and 4).

RL recovers inside a bounded handoff region. The LLaMA mean OOD delta is negative at checkpoints 90–400, positive at every checkpoint from 500–1100, zero at 1200, and negative again from 1300–1600. The public Qwen-DAPO/IFEval sweep in Figure 3a reproduces the same mean-sign



(a) IFEval accuracy (OOD)

(b) AIME-25 accuracy (ID)

Figure 3: Dense Qwen2.5-3B trajectory with OpenR1-Math-220k SFT followed by DAPO. AIME-25 is the ID accuracy metric and IFEval is the OOD accuracy metric. Top panels show SFT and post-DAPO means. Lower panels show the paired post-DAPO minus SFT change; error bars are one standard deviation across three independent training runs. Batch size increases from 16 to 64, so displayed SFT steps equal raw checkpoints divided by four. The IFEval-defined 600–1600 recovery interval is projected onto both panels.

structure: all 11 checkpoints from 600–1600 improve after DAPO, while all nine outer checkpoints decline. The absolute mean change exceeds one seed SD at 10 of 11 inner checkpoints and all nine outer checkpoints. The two dense trajectories therefore identify both an entry and an exit boundary for RL recovery. Evaluation resampling preserves this bounded region, with median entry/exit crossings of 439/1200 for LLaMA/PPO and 447/1121 for Qwen/GRPO (Appendix Table 5). The three-seed robustness sweep preserves negative mean OOD effects at LLaMA checkpoints 300 and 1300 and gains of 3.42–5.13 pp at checkpoint 800 (Appendix Table 6). Base-initialized GeneralPoints RL exhibits unstable loss, while Qwen cold-start GRPO raises AIME-25 avg@16 from 7.70 to 12.29 as GPQA falls from 33.93 to 15.18 and Safety falls from 74.50 to 6.50 (Appendix Figure 16 and Table 10).

Early stopping and late-SFT+RL define distinct operating points. RL from the early $\text{SFT}_{\text{MaxOOD}}$ checkpoint lowers OOD accuracy for both controlled model families, whereas RL from the late handoff restores the forgotten OOD capability. Relative to the stopped early SFT reference, late-SFT+RL retains 37.55 more LLaMA ID points and 20.53 more Qwen ID points while nearly reaching the same OOD accuracy; the three-seed path values appear in Appendix Table 2. ID-only selection follows the continuing controlled-task ID rise and therefore chooses the late handoff rather than the early OOD peak. A validation-based selector retains checkpoints within one seed SD of the best OOD validation score and then selects the candidate with the highest ID validation accuracy (Appendix C.2).

Verifiable rewards support non-unique solutions. Outcome-level rewards reinforce any verifier-correct response, while token-level imitation concentrates supervision on the demonstrated response. This distinction directly favors reasoning tasks with multiple valid derivations. For a single General-Points input, two distinct equations satisfy the verifier, illustrating the multiple derivations available to outcome-level RL (Appendix Figure 20); complete prompts and examples appear in Appendix E.1.

3.3 Evidence Beyond Controlled Tasks

The dense public trajectory trains Qwen2.5-3B with OpenR1-Math-220k SFT and then applies DAPO (Yu et al., 2025). AIME-25 is the ID accuracy metric and IFEval (Zhou et al., 2023) is the OOD accuracy metric. Figure 3 shows the 20-point trajectory, with raw batch-64 checkpoint labels divided by four to match the batch-16 update scale. This dense accuracy experiment is distinct

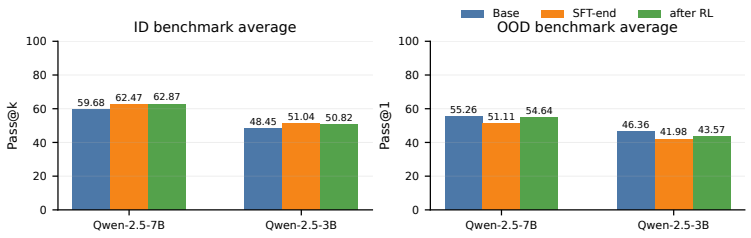


Figure 4: Open-R1 SFT endpoints followed by GRPO for Qwen-2.5-7B and DAPO for Qwen-2.5-3B. Unless noted otherwise, accuracy and rate values in the paper and appendix are reported on a 0–100 scale. ID uses pass@k averages, while OOD uses pass@1 averages.

		ID math/reasoning pass@k							OOD benchmark pass@1						
Model	**Setting**	**AIME24**	**AIME25**	**AMC**	**MATH**	**Minerva**	**Olympiad**	**Avg**	**GPQA**	**IFEval**	**MMLU-Pro**	**SuperGPQA**	**Safety**	**TruthfulQA**	**Avg**
Qwen-2.5-7B	Base	36.67	36.67	90.00	84.60	55.17	54.96	59.68	33.93	61.37	67.14	27.65	74.50	66.96	55.26
Qwen-2.5-7B	SFT-end	50.00	36.67	87.50	87.60	58.09	54.96	62.47	32.80	52.13	65.71	26.42	70.10	59.50	51.11
Qwen-2.5-7B	after GRPO	46.67	33.33	90.00	89.60	58.09	59.55	62.87	33.92	52.31	71.43	32.10	72.90	65.20	54.64
Qwen-2.5-3B	Base	20.00	20.00	80.00	79.20	50.00	41.48	48.45	27.01	54.16	52.86	18.27	65.90	59.94	46.36
Qwen-2.5-3B	SFT-end	30.00	23.33	82.50	80.60	47.42	42.37	51.04	30.58	47.32	45.71	19.26	59.60	49.42	41.98
Qwen-2.5-3B	after DAPO	23.33	33.33	75.00	81.00	47.79	44.44	50.82	27.68	48.43	47.14	22.72	62.50	52.92	43.57

Table 1: Open-R1 SFT followed by GRPO for Qwen-2.5-7B and DAPO for Qwen-2.5-3B. ID columns report pass@k: AIME24/AIME25/AMC use pass@16, MATH-500/Olympiad use pass@4, and Minerva uses pass@8. OOD columns report pass@1. The Base rows provide the pretrained OOD reference.

from the pass@k/pass@1 endpoint evaluation in Table 1. The IFEval curve establishes the bounded recovery region, while the broader endpoint results in Table 1 extend the analysis across task shifts.

We also evaluate endpoints on a broader collection of math and reasoning benchmarks. Figure 4 and Table 1 show that Open-R1 SFT strengthens the ID average while lowering the OOD average, and subsequent RL partially restores the OOD loss. For Qwen-2.5-7B, GRPO recovers 85.1% of the SFT-induced loss in the OOD average: from 55.26 at Base to 51.11 after SFT and to 54.64 after GRPO. For Qwen-2.5-3B, DAPO recovers 36.3%: from 46.36 at Base to 41.98 after SFT and to 43.57 after DAPO. Additional avg@k scoring shows GRPO raising Qwen-2.5-7B ID performance from 37.81 to 43.24 and DAPO raising Qwen-2.5-3B ID performance from 25.65 to 27.73 (Appendix Table 7). Terminal SFT raises controlled-task ID accuracy relative to early SFT in all six model–task pairs; RL then raises Qwen Navigation ID accuracy and lowers the other five endpoints (Appendix Figure 15). Across ten additional model–benchmark OOD pairs, six recover after RL, including full LLaMA recovery on ARC-Challenge and GPQA, while Qwen IFEval rises from 63 to 67 to 70 (Appendix Table 15).

4 Rotation Matters: A SVD Analysis on Parameter Matrices

Section 3.2 establishes the checkpoint-wise behavioral pattern "SFT forgets, RL recovers." We next ask which parameter components carry the functional difference between the forgotten and recovered checkpoints. Guided by spectral analyses of learned weights (Staats et al., 2025; Yunis et al., 2024), we decompose attention and MLP matrices with Singular Value Decomposition (SVD) and directly substitute singular values or singular-vector directions between checkpoints. This design separates changes in coefficient scale from changes in direction and tests which component is more functionally consequential for ID and OOD behavior. Section 4.1 introduces the interventions; Sections 4.2 and 4.3 report the value and direction results.

4.1 Setup

Prior spectral and low-rank studies analyze attention and feed-forward matrices directly (Staats et al., 2025; Wu et al., 2023; Yuan et al., 2024); we therefore study two parameter families:

- **Self-attention projections W_Q, W_K, W_V.** These matrices project hidden states into the query, key, and value subspaces that determine attention scores and contextualized representations (Vaswani et al., 2017).
- **MLP projections W_{MLP}.** Each transformer MLP uses up-, gate-, and down-projections around the SwiGLU nonlinearity (Shazeer, 2020). We analyze the weight matrices; screening experiments found only minor functional effects from the MLP bias term b_{MLP}.

To separate coefficient magnitude from direction, we construct hybrid matrices from checkpoints in the Base–SFT–RL trajectory:

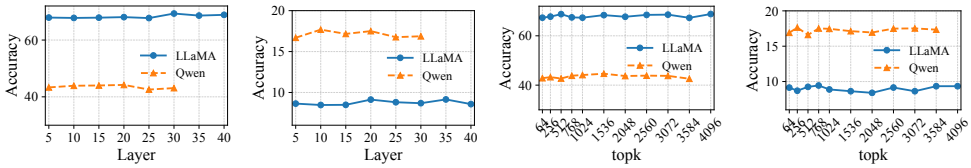
- **Singular-value substitution.** We replace $\Sigma_{\text{SFT}_{\text{End}}}$ with $\Sigma_{\text{SFT}_{\text{MaxOOD}}}$ while retaining SFT_{End} directions, and replace $\Sigma_{\text{RL}_{\text{End}}}$ with $\Sigma_{\text{SFT}_{\text{End}}}$ while retaining RL_{End} directions. The evaluated matrices are $U_{\text{SFT}_{\text{End}}} \Sigma_{\text{SFT}_{\text{MaxOOD}}} V_{\text{SFT}_{\text{End}}}^{\top}$ and $U_{\text{RL}_{\text{End}}} \Sigma_{\text{SFT}_{\text{End}}} V_{\text{RL}_{\text{End}}}^{\top}$.
- **Singular-vector substitution.** We restore checkpoint directions while preserving the destination coefficients, evaluating $U_{\text{SFT}_{\text{MaxOOD}}} \Sigma_{\text{SFT}_{\text{End}}} V_{\text{SFT}_{\text{MaxOOD}}}^{\top}$ and $U_{\text{SFT}_{\text{End}}} \Sigma_{\text{RL}_{\text{End}}} V_{\text{SFT}_{\text{End}}}^{\top}$.

These full-component substitutions diagnose whether behavioral changes follow coefficient scale or direction. Equation 5 provides the targeted intervention: it retains every SFT coefficient and the full SFT tail, replaces only a top-k direction block, and compares the result with five-seed rank-matched random direction sets. For LLaMA, $\text{SFT}_{\text{MaxOOD}} = 140$ and $\text{SFT}_{\text{End}} = 1100$; for Qwen, $\text{SFT}_{\text{MaxOOD}} = 120$ and $\text{SFT}_{\text{End}} = 800$.

To profile how sensitivity accumulates across depth and rank, we apply the substitutions to:

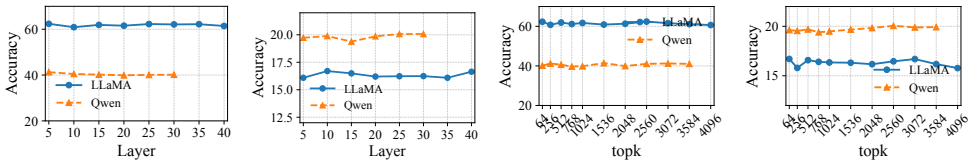
- cumulative layer prefixes $k = 5, 10, 15, 20, \ldots, L$, where L is the number of layers; and
- cumulative top-k singular components with $k = 64, 256, 512, 768, 1024, 1536, 2048, 2560, 3072, 3584$ (and 4096 for LLaMA).

This grid spans small, intermediate, and near-full subspaces and includes the architectural ranks shared by both model families. Sections 4.2 and 4.3 report the results.



(a) Layer-wise (ID) (b) Layer-wise (OOD) (c) Top-k (ID) (d) Top-k (OOD)

Figure 5: Full singular-value diagnostic for the SFT stage on GeneralPoints with LLaMA and Qwen. Across attention and MLP matrices, the panels replace an increasing layer prefix or top-k value block by its $\text{SFT}_{\text{MaxOOD}}$ counterpart while retaining SFT_{End} directions. ID and OOD shifts remain small.



(a) Layer-wise (ID) (b) Layer-wise (OOD) (c) Top-k (ID) (d) Top-k (OOD)

Figure 6: Full singular-value diagnostic for the RL stage on GeneralPoints with LLaMA and Qwen. Across attention and MLP matrices, the panels replace an increasing layer prefix or top-k value block by its SFT_{End} counterpart while retaining RL_{End} directions. ID and OOD shifts remain small.

4.2 Ablation Studies on Singular Values

Prior work often connects singular values to effective capacity, stability, or generalization (Bartlett et al., 2017; Yoshida & Miyato, 2017; Li et al., 2024b). In our controlled substitutions, however,

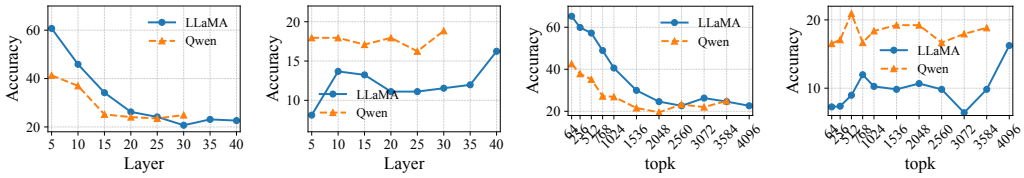
restoring fine-tuned singular values while preserving the checkpoint's singular-vector directions produces only minor ID and OOD changes in both the SFT and RL stages (Figures 5 and 6). Thus, the observed forgetting and recovery cannot be explained by singular-value changes alone.

Appendix G.3 further visualizes low-magnitude, approximately zero-centered value changes for a representative LLaMA attention layer. Singular-value substitutions produce small functional effects across layer prefixes and ranks, whereas the direction substitutions in the next section produce much larger behavioral shifts. This contrast identifies direction identity as a more consequential spectral variable than coefficient magnitude in our setting. The diffuse value changes also differ from settings where spectral updates concentrate in a particular region (Staats et al., 2025; Thamm et al., 2022; Saada et al., 2025; Cancedda, 2024; Hsu et al., 2022).

4.3 Ablation Studies on Singular Vector Directions

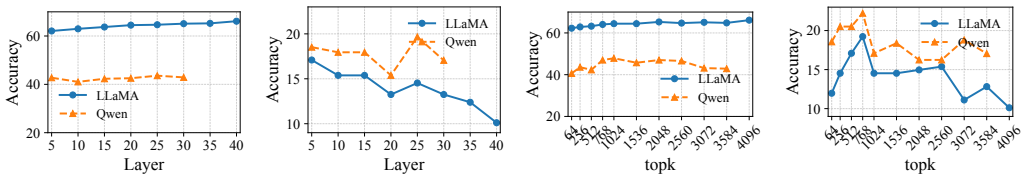
The full singular-vector diagnostics in Figures 7 and 8 produce substantially larger ID and OOD shifts than the value-only substitutions, establishing that model behavior is functionally sensitive to direction identity. An outcome-independent 2×2 natural-scale factorial on stable spectral blocks from 168 Qwen matrices independently restores Base singular values and directions; the equal-six-benchmark OOD macro attributes $+2.81$ pp to directions versus $+0.27$ pp to values, with a $+2.54$ pp direction-minus-value contrast (Appendix G.1). The cumulative layer-prefix and top-k curves identify where adding restored directions changes behavior, while the rank-matched partial intervention in Equation 5 tests direction identity directly with the SFT coefficients and tail held fixed.

In SFT, the cumulative layer-prefix intervention causes large ID degradation once the restored prefix reaches the first 30 LLaMA layers and the first 15 Qwen layers. The LLaMA OOD curve changes sharply when early directions are restored and again when the cumulative intervention reaches the final block, whereas Qwen is less sensitive. The top-k intervention shows the same separation: ID behavior changes strongly by $k = 2560$ for LLaMA and $k = 2048$ for Qwen, while the largest LLaMA OOD recovery appears after restoring dominant directions and again near the full-rank tail. Together, these curves show that OOD forgetting is not carried by singular values alone and that its sensitivity to direction changes spans both dominant and late-added subspaces.



(a) Layer-wise (ID) (b) Layer-wise (OOD) (c) Top-k (ID) (d) Top-k (OOD)

Figure 7: Full singular-vector diagnostic for the SFT stage on GeneralPoints with LLaMA and Qwen. Across attention and MLP matrices, the panels replace directions in an increasing layer prefix or top-k block by $\text{SFT}_{\text{MaxOOD}}$ directions while retaining SFT_{End} coefficients. The resulting ID/OOD shifts are substantially larger than the value-only diagnostic.



(a) Layer-wise (ID) (b) Layer-wise (OOD) (c) Top-k (ID) (d) Top-k (OOD)

Figure 8: Full singular-vector diagnostic for the RL stage on GeneralPoints with LLaMA and Qwen. Across attention and MLP matrices, the panels replace directions in an increasing layer prefix or top-k block by SFT_{End} directions while retaining RL_{End} coefficients. The resulting ID/OOD shifts show broad functional sensitivity to the recovered directions.

During the RL stage, cumulative restoration of SFT_{End} directions reduces LLaMA ID and OOD performance across most layer prefixes, while Qwen is more robust and shows its largest OOD perturbations in intermediate prefixes. The top-k curves show a similar pattern: LLaMA ID performance increases broadly as more SFT directions are restored, and OOD performance is especially sensitive after the dominant block and again near the full-rank tail; Qwen exhibits smaller but directionally consistent changes. These results indicate that the direction-sensitive component of RL recovery is distributed across multiple layers and ranks, rather than attributable to a singular-value shift or a single isolated layer.

Direction-sensitive recovery. In a separate soft subspace-regularized Qwen-2.5-7B/OpenR1-Math comparison, the selected layer-7 W_K right-space coordinate $i = 512$ is 13.929° from Base to standard SFT and 11.791° from Base to subspace-regularized SFT; the subsequent stage-local rotations are 1.025° and 1.028°, respectively (Appendix Table 12). A Qwen2.5-7B dose sweep at target rank 768 further isolates direction identity while matching BF16 displacement and full principal-angle geometry: Base-directed substitution raises IFEval monotonically by $+2.773$ to $+10.536$ pp, while matched random-control means remain between $+0.062$ and $+0.801$ pp; at full dose, the paired gap is $+10.474$ pp (95% CI $[7.209, 13.740]$; Appendix Table 20). The signed activation-projector update points toward Base on all six OOD benchmarks at top-16 and top-32, with mean reversal cosines 0.085 and 0.103 (Appendix Table 16); the all-layer attention K-spectrum correction cosine is 0.146 with a 95% interval of $[0.114, 0.177]$ (Appendix Equation 4). Hidden-state and predictive-distribution distances also contract toward Base on all six benchmarks as OOD pass@1 rises (Appendix Table 17). Most importantly, partial top-k substitution of pretrained directions outperforms five-seed rank-matched random substitutions at every Qwen rank and at LLaMA ranks 512 and 768, reaching 44.231% Qwen and 11.970% LLaMA OOD accuracy at rank 768 (Appendix Table 18). Together, the parameter-, activation-, distribution-, and intervention-level evidence shows that successful recovery contains a consistent Base-directed component, while matched interventions show that restoring pretrained directions can improve OOD performance beyond random direction controls. Detailed definitions and resampling protocols appear in Appendix G.1.

Algorithmic implication. The top-k singular-vector analysis motivates a protected-SFT update that suppresses rotation in OOD-related subspaces. For $W = U\Sigma V^\top$ and SFT gradient G, we project $\widetilde{G}_\perp = (I - U_k U_k^\top) G (I - V_k V_k^\top),\ W \leftarrow W - \eta \widetilde{G}_\perp$, where U_k, V_k are protected top-k singular-vector subspaces. We apply this projection to selected attention and MLP matrices. Appendix Figure 11 and Table 11 show that this preliminary projected-gradient variant preserves more late-SFT OOD accuracy while maintaining comparable ID accuracy. Across three training seeds, the cross-task protected-SFT evaluations raise the OOD macro from 26.67 ± 0.78 to 28.33 ± 0.61 on CommonsenseQA and from 4.71 ± 0.35 to 7.05 ± 2.72 on ProofWriter, while also raising ID accuracy (Appendix Table 13).

5 Conclusions

We refine "SFT memorizes, RL generalizes" into the checkpoint-wise account "SFT forgets, RL recovers." Beyond the finding that terminal RL can improve over terminal SFT, the central finding is that the checkpoint-wise downstream RL effect changes from negative to positive and back to negative as SFT progresses. This entry–recovery–exit geometry is reproduced by dense LLaMA/PPO and Qwen/DAPO trajectories, while three-seed uncertainty and evaluation resampling establish the stability of both forgetting and recovery. The early/late operating-point comparison further shows that late-SFT+RL nearly recovers the early OOD peak while preserving substantially stronger ID capability than oracle early stopping.

The parameter analysis explains why the restored behavior is not captured by singular-value changes alone. Direction substitutions cause large functional shifts, successful RL contains a consistent Base-directed component across parameter and function space, and rank-matched interventions link pretrained direction identity to recovered OOD accuracy. The protected-SFT result turns this analysis into a concrete design implication by preserving more OOD capability during continued supervised training. Together, these results make the SFT handoff checkpoint a first-class component of two-stage post-training and provide a precise account of when RL heals OOD forgetting. Appendix B describes the next scaling and mechanistic extensions.

Reproducibility Statement

Implementation details and experimental settings are described in Sections 3.1 and 4.1 and Appendix E. A code package accompanies this submission as supplementary material.

AI Use Statement

An AI assistant supported editing, reported-data arithmetic checks, and LaTeX formatting.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks, 2017. URL https://arxiv.org/abs/1706.08498.

Åke Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks, 2024. URL https://arxiv.org/abs/2402.09221.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL https://arxiv.org/abs/2501.17161.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jonathan B Freeman and Rick Dale. Assessing bimodality to detect the presence of a dual cognitive process. *Behavior research methods*, 45(1):83–97, 2013.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, 2018.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization, 2022. URL https://arxiv.org/abs/2207.00112.

Jiaji Huang, Qiang Qiu, and Robert Calderbank. The role of principal angles in subspace classification. *IEEE Transactions on Signal Processing*, 64(8):1933–1945, 2015.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. *arXiv preprint arXiv:2507.01679*, 2025.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Aapo Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. *Advances in neural information processing systems*, 10, 1997.

Aapo Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110534, 2013.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead. In *International Conference on Learning Representations*, pp. 56876–56918, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5d6ae8ba43ecb378030753c4408ef9bd-Paper-Conference.pdf.

Minwu Kim, Anubhav Shrestha, Safal Shrestha, Aadim Nepal, and Keith Ross. Reinforcement learning vs. distillation: Understanding accuracy and capability in llm reasoning, 2025. URL https://arxiv.org/abs/2505.14216.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, 2024a.

Yixia Li, Boya Xiong, Guanhua Chen, and Yun Chen. Setar: Out-of-distribution detection with selective low-rank approximation, 2024b. URL https://arxiv.org/abs/2406.12629.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025.

Bingqian Lin, Yunshuang Nie, Khun Loun Zai, Ziming Wei, Mingfei Han, Rongtao Xu, Minzhe Niu, Jianhua Han, Liang Lin, Cewu Lu, et al. Evolvenav: Self-improving embodied reasoning for llm-based vision-language navigation. *arXiv preprint arXiv:2506.01551*, 2025.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Uft: Unifying supervised and reinforcement fine-tuning. *arXiv preprint arXiv:2505.16984*, 2025.

M-A-P Team. SuperGPQA: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025. URL https://arxiv.org/abs/2502.14739.

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

Grant W Petty. On some shortcomings of shannon entropy as a measure of information content in indirect measurements of continuous variables. *Journal of Atmospheric and Oceanic Technology*, 35(5):1011–1021, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

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

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers, 2025. URL https://arxiv.org/abs/2410.07799.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations*, volume 2026, pp. 59839–59864, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/618c95f4557c15b253fb0e6f548ea0c0-Paper-Conference.pdf.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8523–8533, 2025.

Bernard W Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99, 1981.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *Forty-second International Conference on Machine Learning*, 2025.

Max Staats, Matthias Thamm, and Bernd Rosenow. Small singular values matter: A random matrix analysis of transformer models, 2025. URL https://arxiv.org/abs/2410.17770.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. *arXiv preprint arXiv:2506.18880*, 2025.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Physical Review E*, 106(5), November 2022. ISSN 2470-0053. doi: 10.1103/physreve.106.054124. URL http://dx.doi.org/10.1103/PhysRevE.106.054124.

David Tse. Information theory. *Stanford EE/Stats 376A Lecture 15, March 2*, 2017.

Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 10043–10052, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hanyin Wang, Zhenbang Wu, Gururaj Kolar, Hariprasad Korsapati, Brian Bartlett, Bryan Hull, and Jimeng Sun. Reinforcement learning for out-of-distribution reasoning in llms: An empirical study on diagnosis-related group coding. *arXiv preprint arXiv:2505.21908*, 2025.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. Two-stage llm fine-tuning with less specialization and more generalization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024b.

Yifan Wu, Shichao Kan, Min Zeng, and Min Li. Singularformer: Learning to decompose self-attention to linearize the complexity of transformer. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 4433–4441. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/493. URL https://doi.org/10.24963/ijcai.2023/493. Main Track.

Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of SFT: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, pp. 27550–27571, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2e53c02ea028cbf603f4b6b47fef3d97-Paper-Conference.pdf.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2502.14768.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning, 2017. URL https://arxiv.org/abs/1705.10941.

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

Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models, 2024. URL https://arxiv.org/abs/2312.05821.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, pp. 57654–57689, 2025. doi: 10.52202/

085713-1933. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/537d5aa768c2d534016a4d06f87bc8fb-Paper-Conference.pdf.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R Walter. Approaching deep learning through the spectral dynamics of weights. *arXiv preprint arXiv:2408.11804*, 2024.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in neural information processing systems*, 37:110935–110971, 2024.

Miaosen Zhang, Yishan Liu, Shuxia Lin, Xu Yang, Qi Dai, Chong Luo, Weihao Jiang, Peng Hou, Anxiang Zeng, Xin Geng, and Baining Guo. Towards on-policy SFT: Distribution discriminant theory and its applications in LLM training, 2026. URL https://arxiv.org/abs/2602.12222.

Anhao Zhao, Ziyang Chen, Junlong Tong, Yingqi Fan, Fanghua Ye, Shuhao Li, Yunpu Ma, Wenjie Li, and Xiaoyu Shen. On-policy supervised fine-tuning for efficient reasoning, 2026. URL https://arxiv.org/abs/2602.13407.

Yilun Zheng, Sha Li, Fangkun Wu, Yang Ziyi, Lin Hongchao, Zhichao Hu, Cai Xinjun, Ziming Wang, Jinxuan Chen, Sitao Luan, Jiahao Xu, and Lihui Chen. FanChuan: A multilingual and graph-structured benchmark for parody detection and analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21937–21957. Association for Computational Linguistics, July 2025. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1131. URL https://aclanthology.org/2025.findings-acl.1131/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. Anchored supervised fine-tuning. In *International Conference on Learning Representations*, pp. 110364–110385, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b2f57eb82f9ef6bb080890ba33a81760-Paper-Conference.pdf.

Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

APPENDIX CONTENTS

A Related Work **16**

B Limitations **17**

C Additional Evidence and Stability Analyses **17**

C.1 Early and Late Handoff Operating Points and Seed Stability 17

C.2 Checkpoint Selection 18

C.3 LLaMA Recovery across RL Budgets 18

C.4 Broader Benchmark Details: Open-R1 SFT Followed by RL 18

C.5 Boundary Evidence Beyond the Main LLaMA Sweep 19

C.6 Cold-Start RL Evidence 21

C.7 Mitigation and Singular-Vector Evidence 21

D Operational Distinctions: Forgetting, Specialization, Over-Fitting, and Over-Training **22**

E More Details of Experimental Settings **23**

E.1 Task Descriptions and Prompts 23

E.2 Computational Resources and Setups 25

F More Experimental Results **25**

F.1 ID and OOD Loss in SFT 25

F.2 Results with GRPO 26

F.3 More Results of In-Distribution Generalization Performance 26

F.4 OOD Results on Other Benchmark Tasks 27

F.5 Loss of Single-Stage RL Fine-tuning 27

G Additional SVD Analysis **28**

G.1 Direction-Sensitive Recovery and Rank-Matched Controls 28

G.2 SVD Notation 31

G.3 Changes of Singular Values 31

G.4 Exploring the Rotation of Singular Vector with Principal Angles 31

G.5 PCA Visualization of Embedding Shifts 34

G.6 Early Verifier-Error Decomposition 34

G.7 Advantage Distribution Metrics 35

G.8 Full Results of Advantage Distribution 37

A Related Work

SFT-Side Generalization and On-Policy Supervision. Recent work improves generalization during supervised training. Wu et al. (2026) express the SFT gradient in policy-gradient form and introduce Dynamic Fine-Tuning (DFT) to correct inverse-probability distortion through probability-weighted token updates. Zhu et al. (2026) place SFT and DFT in a reward-weighted regression framework and add KL anchoring to the pretrained reference. Zhao et al. (2026) reduce a correctness–efficiency RL objective to supervised learning on self-generated responses filtered for correctness and conciseness, while Zhang et al. (2026) adjust model–data alignment through IDFT and Hinted Decoding. Ren et al. (2026) further show that generalization in reasoning SFT can be non-monotonic across optimization, data, and model-capability conditions. We examine these dynamics across the SFT–RL pipeline by locating the early OOD peak, tracking the downstream RL effect across handoff checkpoints, and comparing early stopping with late-SFT+RL.

RL Improves Reasoning and OOD Generalization. Large-scale RL fine-tuning has substantially improved LLM performance in mathematics and domain-specific reasoning (DeepSeek-AI, 2025; Wang et al., 2025). These advances have motivated a growing body of work on the differences between RL and SFT in reasoning and generalization. In particular, Chu et al. (2025) argue that SFT tends to memorize training distributions, whereas RL can induce more transferable reasoning behavior. Rule-based RL also improves reasoning on synthetic logic tasks and transfers to harder benchmarks such as AIME and AMC (Xie et al., 2025). Other work integrates SFT and RL by incorporating supervised signals into RL-style training objectives (Liu et al., 2025; Huang et al., 2025). Together, these studies highlight the role of RL in modern reasoning post-training.

SFT-to-RL Handoff. Kang et al. (2026) evaluate independently trained SFT candidates at scale and show that terminal SFT accuracy can misrank initializers for subsequent RL. Held-out generalization loss and pass@large-k are stronger predictors of final post-RL performance. We trace the paired downstream OOD effect of RL across checkpoints along a single SFT trajectory. The effect is negative for early initializers, positive within an intermediate window, and negative again after the exit boundary. The resulting entry and exit boundaries connect initializer selection to OOD capability first exposed and then forgotten during SFT.

Base-Relative Retention. Shenfeld et al. (2026) show that on-policy RL is implicitly biased toward KL-minimal solutions near the starting policy and therefore preserves prior capabilities better than SFT at matched new-task performance. Their result characterizes the retention bias of online RL. Our trajectory analysis tracks OOD recovery as SFT progresses, identifying an intermediate recovery window between early checkpoints with limited task competence and late checkpoints beyond the recovery boundary. Our Base-relative activation, distributional, and spectral analyses, together with direction interventions, connect OOD recovery to specific pretrained subspaces.

Does RL Create New Reasoning Capability? Several studies probe the capability boundaries of RL-trained models. Yue et al. (2025) analyze RLVR from a fixed initializer and show that high-k sampling exposes correct reasoning paths already present in that policy, while RLVR improves low-k sampling efficiency. Kim et al. (2025) find that RL can favor easier problems over harder ones, while Zheng et al. (2025) find that reasoning LLMs generally underperform non-reasoning LLMs on parody detection, illustrating uneven generalization beyond math and code. Our experiments vary the initializer along SFT and measure the signed downstream OOD effect of RL, identifying the checkpoints where recovery begins and ends. This trajectory-level analysis complements the study of reasoning support and sampling efficiency within a fixed initializer.

Our Perspective. Chu et al. (2025) compare continued SFT with PPO from one SFT checkpoint and establish the endpoint contrast summarized as "SFT memorizes, RL generalizes." Our checkpoint-wise analysis traces the dynamics behind this contrast: OOD capability appears early, later SFT erodes it while ID performance improves, and the downstream RL effect changes sign twice along the SFT trajectory. The dense public-benchmark trajectory reproduces the bounded mean-sign recovery window, while broader evaluations extend the analysis across cross-task shifts, model sizes, datasets, and RL algorithms. Together, the three-seed trajectories, early/late operating-point comparison, and

Base-relative interventions provide a checkpoint-wise account of post-training: "SFT forgets, RL recovers."

B LIMITATIONS

Our empirical scope combines controlled rule shifts with public-benchmark trajectories and multi-benchmark endpoints, exposing checkpoint dynamics across explicit rule changes and cross-task shifts. Scaling the dense protocol to larger model families offers the next empirical extension.

C ADDITIONAL EVIDENCE AND STABILITY ANALYSES

This section collects auxiliary evidence for the main claims: broader reasoning benchmarks and DAPO results, boundary sweeps, cold-start RL evidence, and mitigation/mechanistic evidence.

C.1 EARLY AND LATE HANDOFF OPERATING POINTS AND SEED STABILITY

Table 2 compares the early Max-OOD handoff with the terminal SFT handoff. RL from the early checkpoint lowers OOD accuracy in both model families, whereas RL from the late checkpoint restores the forgotten capability. The stopped early SFT model is also reported as the oracle trajectory reference. Relative to that reference, late-SFT+RL retains 37.55 additional LLaMA ID points and 20.53 additional Qwen ID points while reaching nearly the same OOD accuracy.

Model / RL / task	Stage	Checkpoint	ID acc.	OOD acc.
LLaMA-3.2-11B / PPO / GeneralPoints	SFT Max-OOD	140	24.790 ± 2.820	17.520 ± 2.290
	RL from Max-OOD	140 $\rightarrow$RL	15.810 ± 2.380	8.550 ± 1.830
	SFT end	1100	68.380 ± 3.242	6.920 ± 1.529
	RL from SFT end	1100 $\rightarrow$RL	62.340 ± 3.170	16.240 ± 2.410
Qwen-2.5-7B / GRPO / GeneralPoints	SFT Max-OOD	120	26.070 ± 2.870	19.670 ± 2.600
	RL from early SFT	120 $\rightarrow$RL	24.590 ± 3.220	16.470 ± 2.820
	SFT end	800	43.590 ± 3.240	14.100 ± 2.460
	RL from SFT end	800 $\rightarrow$RL	46.600 ± 3.210	19.660 ± 2.600

Table 2: Early- and late-handoff operating points. Accuracy is mean $\pm$ one SD over three independent training runs.

The LLaMA seed-level effect sizes and evaluation-resampling intervals in Table 3 separate both forgetting and recovery from zero. Peak-location resampling in Table 4 concentrates the LLaMA peak in checkpoints 90–300 and the Qwen peak in the early 110–400 region. The Qwen peak-localization analysis uses a finer early grid that includes checkpoint 120; the extended boundary table uses a 100-step display grid.

Setting	Quantity	Estimate	Seed SD	95% resampling interval
LLaMA-3.2-11B / PPO / GeneralPoints	Forgetting magnitude	10.600 pp	2.990 pp	[4.700, 16.110]
LLaMA-3.2-11B / PPO / GeneralPoints	RL recovery magnitude	9.320 pp	3.377 pp	[3.620, 14.210]
Qwen-2.5-7B / GRPO / GeneralPoints	Forgetting magnitude	5.570 pp	—	—
Qwen-2.5-7B / GRPO / GeneralPoints	RL recovery magnitude	5.560 pp	—	—

Table 3: Stability of the forgetting and recovery magnitudes. Point estimates are differences of three-run mean accuracies in Table 2. LLaMA SDs use paired training runs; intervals resample evaluation examples over their outputs. Qwen rows report mean contrasts only.

Setting	Observed peak	Exact-peak frequency	Peak-region frequency	Central 90% range
LLaMA-3.2-11B / PPO / GeneralPoints	140	49.1% at 140	98.8% in 90–300	140–300
Qwen-2.5-7B / GRPO / GeneralPoints	120	37.7% at 120	59.3% in 120–300	110–400

Table 4: Evaluation-resampled stability of the Max-OOD checkpoint location.

C.2 Checkpoint Selection

For each evaluation resample, we linearly interpolate the paired RL effect across adjacent SFT checkpoints. Table 5 reports the median resampled crossing as the point estimate and the central 95% resampling range; both settings retain the entry/exit structure.

Setting	Boundary	Estimate	95% sensitivity range
LLaMA-3.2-11B / PPO / GeneralPoints	Entry	439	[372, 813]
LLaMA-3.2-11B / PPO / GeneralPoints	Exit	1200	[1162, 1325]
Qwen-2.5-7B / GRPO / GeneralPoints	Entry	447	[311, 746]
Qwen-2.5-7B / GRPO / GeneralPoints	Exit	1121	[871, 1179]

Table 5: Median resampled entry and exit crossings with central 95% evaluation-resampling sensitivity ranges.

A practical selector uses a predefined checkpoint grid and a held-out OOD validation set. First retain checkpoints within one seed SD of the best OOD validation score, then choose the checkpoint with the highest ID validation accuracy, and finally evaluate that checkpoint once on test. For an RL handoff, generate a fixed set of ID responses with the unchanged decoder and report the binary-reward advantage mean, SD, and positive/negative fractions. Substantial mass on both sides of zero supplies simultaneous reinforcement and correction signals. This validation rule converts the observed trajectory geometry into a setting-calibrated selection procedure while preserving a single-use test set.

C.3 LLaMA Recovery across RL Budgets

We vary the PPO training budget for LLaMA-3.2-11B-Vision on GeneralPoints at SFT checkpoints 300, 800, and 1300, sampling the region before entry, the interior of the recovery window, and the region after exit, respectively. The four budgets are $0.5\times$, $1\times$, $1.5\times$, and $2\times$, relative to the PPO run used in Figure 2c. Each setting uses three independent training seeds. Table 6 combines the additional runs with the SFT and $1\times$ results from that trajectory.

		Post-PPO OOD accuracy (%)			
SFT checkpoint	SFT OOD (%)	$0.5\times$	$1\times$	$1.5\times$	$2\times$
300	16.24 ± 2.58	7.26 ± 1.20	8.12 ± 1.54	9.12 ± 2.58	9.54 ± 1.62
800	5.98 ± 1.08	11.11 ± 0.12	9.40 ± 1.67	10.26 ± 0.43	11.11 ± 1.28
1300	9.40 ± 2.21	6.41 ± 1.44	6.84 ± 1.51	6.84 ± 1.86	6.70 ± 1.93

Table 6: LLaMA/PPO/GeneralPoints OOD accuracy across four RL budgets. Values are mean $\pm$ one standard deviation over three training seeds, rounded to two decimal places. The SFT and $1\times$ columns reproduce the corresponding points in Figure 2c. The mean RL effect retains the negative–positive–negative pattern at checkpoints 300, 800, and 1300 under every tested budget.

At checkpoint 300, post-PPO mean OOD accuracy remains 6.70–8.98 pp below SFT. At checkpoint 800, it improves by 3.42–5.13 pp, while at checkpoint 1300 it remains 2.56–2.99 pp below SFT. Thus, the interior recovery and the negative effects on both sides persist across a fourfold range of RL budgets at these tested checkpoints.

C.4 Broader Benchmark Details: Open-R1 SFT Followed by RL

The main paper reports the broader benchmark summary in Table 1, including all ID pass@k scores, all OOD pass@1 scores, and their averages. To avoid repeating the same results, this appendix reports only the additional ID avg@k datapoints for Open-R1 SFT (Hugging Face, 2025) followed by GRPO on Qwen-2.5-7B and DAPO (Yu et al., 2025) on Qwen-2.5-3B. These results complement

the pass@k scores in the main paper: both RL stages improve the average ID reasoning score after SFT, while the main-paper OOD pass@1 table shows partial OOD restoration.

Model	Setting	AIME24 avg@16	AIME25 avg@16	AMC avg@16	MATH-500 avg@4	Minerva avg@8	Olympiad avg@4	Avg(avg@k)
Qwen-2.5-7B	Base	12.08	7.70	52.03	72.85	37.09	37.41	36.53
Qwen-2.5-7B	SFT-end	15.21	19.38	50.47	72.85	31.52	37.41	37.81
Qwen-2.5-7B	after GRPO	18.75	20.83	58.13	79.35	38.97	43.44	43.24
Qwen-2.5-3B	Base	4.79	2.29	37.03	63.35	26.24	26.22	26.65
Qwen-2.5-3B	SFT-end	3.96	5.00	35.00	60.95	22.75	26.25	25.65
Qwen-2.5-3B	after DAPO	4.79	5.42	37.50	65.90	25.32	27.44	27.73

Table 7: Additional ID avg@k results for Open-R1 SFT followed by GRPO on Qwen-2.5-7B and DAPO on Qwen-2.5-3B. The corresponding ID pass@k results and all OOD pass@1 results are reported in the main paper in Table 1.

C.5 Boundary Evidence Beyond the Main LLaMA Sweep

We further test the recovery boundary with two complementary datasets. Figure 9 summarizes the recovery deltas, Figure 10 and Table 8 give the full Qwen checkpoint sweep with GRPO on GeneralPoints, and Table 9 gives the sparse boundary evaluation for a broader reasoning setup using Open-R1 SFT followed by DAPO-math RL. Exact boundary locations vary by benchmark; the shared structure is an intermediate positive region bracketed by negative outer checkpoints.

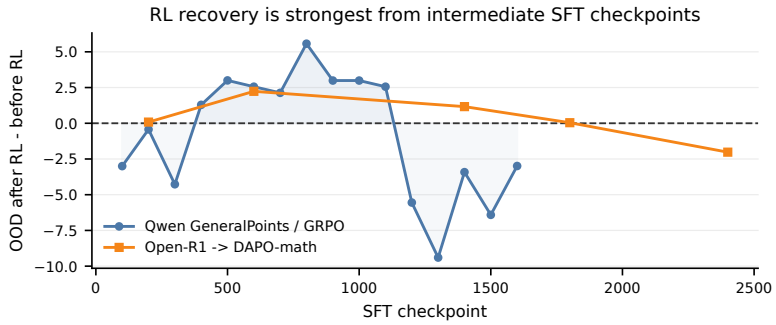


Figure 9: OOD recovery deltas for Qwen with GRPO and Open-R1 SFT followed by DAPO-math RL. Positive values mean RL improves OOD over the SFT initialization.

SFT checkpoint	ID before RL	OOD before RL	ID after RL	OOD after RL
100	20.09	19.67	20.59	16.67
200	25.21	17.52	25.84	17.09
300	32.05	17.52	32.85	13.25
400	32.48	15.38	33.29	16.67
500	32.05	15.38	32.85	18.38
600	35.04	16.24	35.90	18.80
700	32.91	16.67	34.19	18.80
800	43.59	14.10	47.74	19.67
900	47.86	11.97	49.06	14.96
1000	49.57	13.68	50.81	16.67
1100	57.26	13.68	63.25	16.24
1200	61.11	17.52	60.26	11.97
1300	60.70	17.95	57.26	8.55
1400	64.10	13.25	62.82	9.83
1500	61.11	14.10	67.52	7.69
1600	67.52	12.82	73.08	9.83

Table 8: Full Qwen checkpoint sweep with GRPO on GeneralPoints. The OOD recovery effect is strongest for intermediate checkpoints and weakens for late checkpoints.

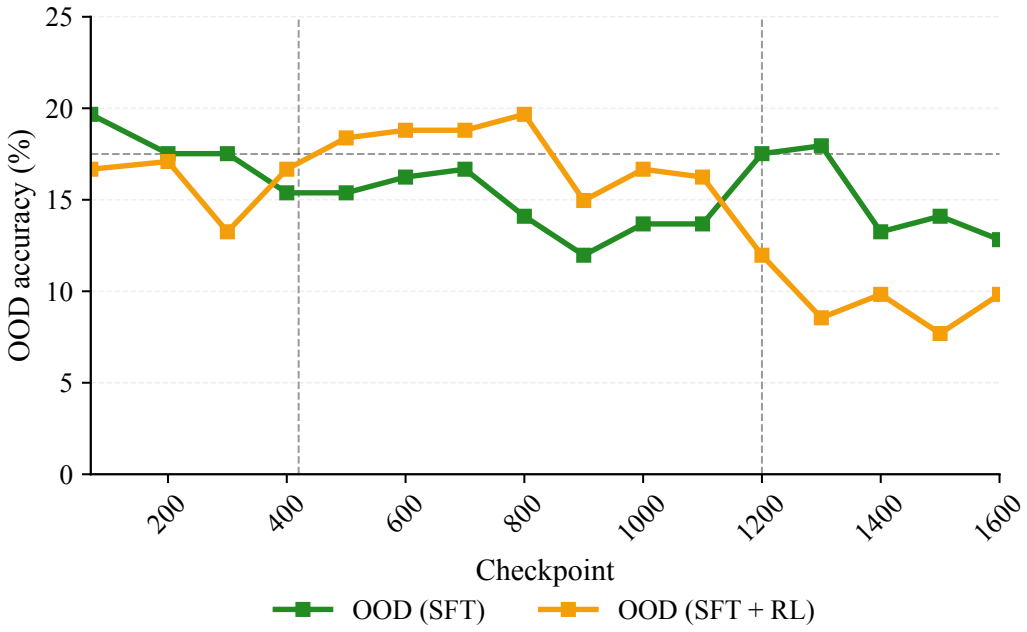


Figure 10: Qwen checkpoint-wise OOD recovery boundary with GRPO on GeneralPoints.

Checkpoint	Setting	IFEval-loose	MMLU-Pro	Safety	TruthfulQA-MC
0	Base	54.16	52.86	65.90	59.94
200	SFT-end	51.76	44.29	62.70	54.82
200	after DAPO	49.91	47.14	62.90	53.95
600	SFT-end	47.32	45.71	59.60	49.42
600	after DAPO	48.43	47.14	62.50	52.92
1400	SFT-end	44.54	45.71	59.50	48.83
1400	after DAPO	44.92	47.14	60.30	50.88
1800	SFT-end	42.88	41.43	62.80	49.42
1800	after DAPO	41.96	44.28	60.60	49.85
2400	SFT-end	39.37	42.86	61.80	50.00
2400	after DAPO	38.08	42.85	58.00	47.01

Table 9: Sparse Qwen-2.5-3B boundary evaluation with Open-R1 SFT followed by DAPO-math RL. Intermediate checkpoints improve on most OOD benchmarks, while the earliest and latest checkpoints show mixed or negative effects.

C.6 Cold-Start RL Evidence

Table 10 reports the Qwen-2.5-7B cold-start GRPO result. Target-math scores rise, while GPQA, IFEval, and Safety fall. This ID–OOD trade-off supports the entry-boundary interpretation: SFT task competence changes which behaviors the subsequent RL stage can reinforce.

Setting	GPQA	IFEval-loose	Safety	AIME24 avg@16	AIME25 avg@16	AMC avg@16
Base	33.93	61.37	74.50	12.08	7.70	52.03
Cold-start GRPO	15.18	58.96	6.50	13.33	12.29	53.13

Table 10: Qwen-2.5-7B cold-start RL with GRPO.

C.7 Mitigation and Singular-Vector Evidence

Figure 11 and Table 11 provide the single-run projected-gradient protected-SFT trajectory on GeneralPoints. The projected-gradient protection raises OOD accuracy at 9 of 12 checkpoints and at 4 of the 5 late checkpoints from 800–1200, while late ID accuracy remains comparable.

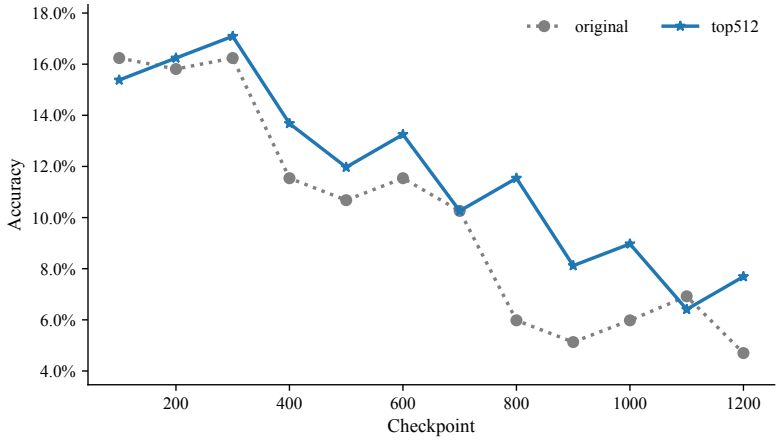


Figure 11: Single-run GeneralPoints protected-SFT trajectory using the top-k projected-gradient update on selected attention and MLP matrices. The singular-vector protection raises OOD accuracy at 9 of 12 checkpoints while retaining comparable late-stage ID accuracy.

Checkpoint	OOD	OOD protected	ID	ID protected
100	16.24	15.38	24.79	21.79
200	15.81	16.24	31.62	26.07
300	16.24	17.09	37.18	34.19
400	11.54	13.68	37.61	36.19
500	10.68	11.97	45.73	44.87
600	11.54	13.25	50.00	47.86
700	10.26	10.26	58.12	60.68
800	5.98	11.54	64.00	66.67
900	5.13	8.12	70.09	69.66
1000	5.98	8.97	66.67	70.35
1100	6.92	6.41	68.38	72.65
1200	4.70	7.69	76.07	75.68

Table 11: Single-run SFT with top-512 singular-vector protection on GeneralPoints. Protection preserves more OOD accuracy at 4 of the 5 late checkpoints from 800–1200 while maintaining comparable ID accuracy.

angle index i	Base→SFT	Base→Subspace-reg. SFT	SFT→RL	Subspace-reg. SFT→RL
1	0.0000	0.0000	0.274112	0.464360
2	0.0000	0.0000	0.472712	0.480513
4	0.0000	0.0000	0.596433	0.562322
8	0.0000	0.0000	0.681275	0.707206
16	0.0885	0.0000	0.767958	0.787830
32	0.6904	0.0000	0.920899	0.918132
64	0.7130	0.0000	0.938786	0.941908
128	0.7538	0.0000	0.956544	0.958383
256	0.8379	0.0000	0.977187	0.979388
384	0.8838	0.7044	1.000342	1.003272
512	13.9290	11.7914	1.025075	1.027744

Table 12: Selected ordered right-subspace canonical angles for the fixed top-512 comparison of Qwen-2.5-7B layer-7 W_K after OpenR1-Math training. Each row reports the individual angle coordinate θ_i among the 512 ordered angles, rather than a top-i subspace aggregate or a mean across layers or matrix families. The first two columns compare Base with the standard-SFT and soft subspace-regularized-SFT endpoints; the final two compare each SFT endpoint with its subsequent RL endpoint. Thus, 13.9290° is θ_{512} for this layer-7 matrix. The soft block-drift regularizer used here is distinct from the projected-gradient protected-SFT experiments in Figure 11 and Table 11.

Table 12 reports the separate soft subspace-regularized Qwen-2.5-7B/OpenR1-Math layer-7 fixed-top-512 comparison. Its sharp tail value is $\theta_{512} = 13.9290^\circ$. Figure 18 displays the full 25–30° to 90° ordered-angle spectrum for a separate layer-0 W_K matrix. Both results contain per-matrix angle coordinates. The 28-layer aggregation appears only in the key-spectrum correction in Equation 4.

We further test full-parameter protected SFT with Qwen-2.5-7B on the `tau/commonsense_qa` release of CommonsenseQA (Talmor et al., 2019)[1] and the `tasksource/proofwriter` release of ProofWriter (Tafjord et al., 2021)[2]. The protected rank is $k = 256$, and Table 13 reports the matched one-epoch checkpoint over three independent training seeds. OOD evaluation uses IFEval (Zhou et al., 2023), MMLU-Pro (Wang et al., 2024b), and SuperGPQA (M-A-P Team, 2025). Protection raises both source-task ID accuracy and the OOD macro: $84.30 \pm 0.90 \rightarrow 84.58 \pm 0.21$ ID and $26.67 \pm 0.78 \rightarrow 28.33 \pm 0.61$ OOD macro for CommonsenseQA, and $89.42 \pm 1.32 \rightarrow 91.16 \pm 1.54$ ID and $4.71 \pm 0.35 \rightarrow 7.05 \pm 2.72$ OOD macro for ProofWriter.

SFT dataset	Method	ID	IFEval	MMLU-Pro	SuperGPQA	OOD macro
CommonsenseQA	Base	79.69 ± 0.00	61.92 ± 0.00	60.00 ± 0.00	16.79 ± 0.00	46.24 ± 0.00
	SFT (epoch 1.0)	84.30 ± 0.90	24.58 ± 1.33	40.95 ± 2.96	14.49 ± 0.74	26.67 ± 0.78
	Protected SFT (epoch 1.0)	84.58 ± 0.21	27.73 ± 0.64	42.38 ± 0.81	14.90 ± 0.57	28.33 ± 0.61
ProofWriter	Base	72.51 ± 0.00	61.92 ± 0.00	60.00 ± 0.00	16.79 ± 0.00	46.24 ± 0.00
	SFT (epoch 1.0)	89.42 ± 1.32	14.05 ± 0.95	0.00 ± 0.00	0.08 ± 0.14	4.71 ± 0.35
	Protected SFT (epoch 1.0)	91.16 ± 1.54	14.85 ± 1.92	5.71 ± 7.55	0.58 ± 0.57	7.05 ± 2.72

Table 13: Three-seed protected-SFT transfer on CommonsenseQA and ProofWriter. Values are accuracy (%) reported as mean ± one SD over three independent training seeds. Base is the shared pretrained checkpoint and therefore has zero seed variation. Its OOD values repeat because both blocks use the same frozen model and shared OOD evaluations, while Base ID differs across the two held-out source datasets. ID is source-dataset accuracy; OOD macro is the unweighted mean of IFEval, MMLU-Pro, and SuperGPQA accuracy.

D Operational Distinctions: Forgetting, Specialization, Over-Fitting, and Over-Training

We use the following operational distinctions throughout the paper.

[1] `https://huggingface.co/datasets/tau/commonsense_qa`

[2] `https://huggingface.co/datasets/tasksource/proofwriter`

- **Catastrophic forgetting** is the loss of previously acquired knowledge or skills during adaptation to new data (Li et al., 2024a; Kotha et al., 2024).

- **Over-specialization** is an increasingly narrow fit to the supervised task's format or rule (Wang et al., 2024a). In our controlled setting, later SFT specializes to the ID face-card rule and reduces shifted-rule accuracy while held-out ID accuracy rises.

- **Over-fitting** is a widening train–held-out gap under a fixed evaluation distribution. Standard ID loss and accuracy monitor this within-distribution effect; our OOD-forgetting metric instead compares accuracy across an explicit rule or task shift.

- **Over-training** describes a training-duration effect in which further optimization degrades a downstream objective. Pretraining over-training can also increase post-training sensitivity (Springer et al., 2025).

Thus, **OOD forgetting** in this paper denotes the measured decline from an earlier OOD checkpoint to a later SFT checkpoint under a fixed evaluation protocol, including cases where ID validation performance continues to rise.

E More Details of Experimental Settings

E.1 Task Descriptions and Prompts

GeneralPoints Prompts and examples for the GeneralPoints game are shown as follows.

[Task Description]
You are an expert 24 points card game player. You will receive a set of 4 cards. Face-card values follow the split-specific rule stated in the prompt: the ID rule maps 'J', 'Q', and 'K' to 10, while the OOD rule maps them to 11, 12, and 13. Each card must be used once. Your goal is to output a formula that evaluates to 24 using numbers from the cards and operators such as '+', '-', '*', '/', and '='.
[Input]
Cards: [1, 3, K, 6]
[Output]
{ "cards": [x, y, z, w], where {face_card_msg},
"number": [a, b, c, d], where a, b, c, and d are the numbers on the cards,
"formula": 'an equation that equals 24',
}
For In-distribution Response:
{
"cards":[1, 3, K, 6],
"number":[1, 3, 10, 6],
"formula":"$(10 \times 3) - (6 \div 1) = 24$"
}

For Out-of-Distribution Response:
{
"cards":[1, 3, K, 6],
"number":[1, 3, 13, 6],
"formula":"$(6 \times (13 - 1)) \div 3 = 24$"
}

Navigation For the Navigation task, which is used in (Chu et al., 2025), we train the model with an absolute direction *e.g.,* turn (source west), then we evaluate the model's OOD performance by relative direction *e.g.,* turn(left). Prompts and examples for Navigation are shown as follows.

[Task Description]
You are an expert in navigation. You will receive a sequence of instructions to follow. You are also provided with your observation and action history in text. Your goal is to first analyze the instruction and identify the next sentence to be executed. Then, you need to provide the action to be taken based on the current observation and instruction.
[Instruction]
1. First, turn right to face north.
2. Move forward until you reach next intersection.
3. Turn left to face west.
4. Move forward until you reach next intersection.
5. Turn left to face north.
6. Move forward until you reach next intersection.
7. Turn right to face east.
8. Move forward until you reach next intersection where Levi & Korsinsky, LLP is on your right behind.
9. Turn left to face north.
10. Move forward until you reach next intersection.
11. Turn slightly right to face northeast.
12. Move forward until you reach next intersection.
13. Turn right to face northwest.
14. Move forward until you reach next intersection where Mr Goods Buy & Sell is on your left front.
15. Turn left to face northeast.
16. Move forward until you reach next intersection where Skullfade Barbers is on your left front.
17. Turn right to face northwest.
18. Move forward until you reach destination where The destination Ann Cleaners is on your left.

[Action space]
forward(): indicates moving forward one step
turn direction(x): adjusts the ego direction. ID prompts use one of the eight absolute compass directions ['north', 'northeast', 'east', 'southeast', 'south', 'southwest', 'west', 'northwest']; OOD prompts use relative directions such as 'left' and 'right'.
stop(): indicates the navigation is finished.
vspace6pt

[Observations and action sequence]
O_1: No landmarks nearby;
A_1:
For In-distribution Response:
{
"current observation": "No landmarks nearby; "
"current instruction": "First, turn right to face north."
"action": "turn direction(north)"
}

For Out-of-Distribution Response:
{
"current observation": "No landmarks nearby; "
"current instruction": "First, turn right to face north."
"action": "turn direction (right)"
}

Rank-Determinant Computation For the matrix computation task, we train LLaMA and Qwen to compute the rank of a matrix with a given dimension, *e.g.,*, 4×5, then evaluate determinant computation as the OOD task, adapted from (Sun et al., 2025). The shift evaluates both arithmetic and cross-concept linear-algebra transfer.

For in-distribution training, the prompt is:

[Task Description]
You are an expert in linear algebra. You will receive a matrix. Find the rank of the matrix and output the integer result.
[Input]
Matrix:[[-1, -2, 9, 3, -5], [0, -3, 9, 9, -6], [-2, -2, 12, 0, -6], [3, -2, -3, 15, -1]]
[Output]
{ "answer": 2,
}

For out-of-distribution evaluation, the prompt is:

[Task Description]
You are an expert in linear algebra. You will receive a square matrix. Compute its determinant and output the integer result.
[Input]
Matrix:[[-4, 3], [-3, -2]]
[Output]
{ "answer": 17,
}

E.2 Computational Resources and Setups

Controlled jobs use four to eight GPUs according to model and algorithm. Controlled SFT uses four GPUs, a learning rate of 10^{-6}, mini-batch size 64, and cosine learning-rate decay. Controlled PPO/GRPO uses a 10^{-6} learning rate and temperature 0.2. Controlled evaluation covers 234 examples under the deterministic task verifier, with decoding held fixed inside each comparison. The dense LLaMA/PPO/GeneralPoints sweep covers 17 SFT initializers from checkpoint 90 to 1600. The public Qwen2.5-3B/OpenR1-Math/DAPO sweep covers 20 initializers from normalized step 100 to 2000 with batch size 64; normalized labels equal raw checkpoint divided by four. The remaining model/task combinations use targeted endpoint or boundary sweeps summarized in Table 14.

The standard controlled endpoints are LLaMA checkpoint 1100 and Qwen checkpoint 800; the extended boundary sweeps continue to checkpoint 1600. Public-math RL rollouts use temperature 1 and top-$p = 1$. Pass@k sample counts appear in Table 1.

Table 14: Details of implementation configurations.

Task	Model	MaxOOD	SFT data	RL begin	Eval data
GeneralPoints	LLaMA	140	100k	500–1100	234
GeneralPoints	Qwen	120	80k	400–800	234
Navigation	LLaMA	45	5k	60	234
Navigation	Qwen	95	10k	100	234
Matrix	LLaMA	50	15k	140	234
Matrix	Qwen	650	80k	800	234

The task prompts and evaluation protocols appear in the surrounding appendix sections.

F More Experimental Results

F.1 ID and OOD Loss in SFT

After checkpoint 50, ID cross-entropy approaches 0.15 while OOD loss rises; OOD accuracy nevertheless continues improving through checkpoint 140 in Figure 2c. This **loss–accuracy discrepancy** appears for both LLaMA and Qwen. Format alignment completes near checkpoint 50, after which specialization to the ID card rule increases number-interpretation errors for OOD face cards. At the same time, arithmetic reasoning continues improving, raising the probability of a correct formula whenever the numbers are interpreted correctly. Token-level loss aggregates both effects, whereas verifier accuracy measures end-to-end task success; token-level cross-entropy therefore incompletely reflects reasoning capacity.

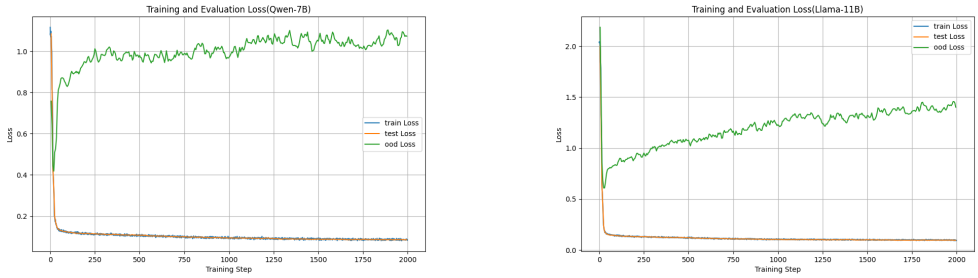


Figure 12: In-distribution training/test loss and OOD loss curves for LLaMA-3.2-11B-Vision and Qwen-2.5-7B during SFT.

F.2 Results with GRPO

We verify the trajectory with GRPO (group size $= 4$); Figures 13 and 14 report LLaMA and Qwen results on GeneralPoints. GRPO yields stronger ID and lower OOD than PPO, while both algorithms heal late-SFT OOD forgetting and approach the early SFT OOD peak.

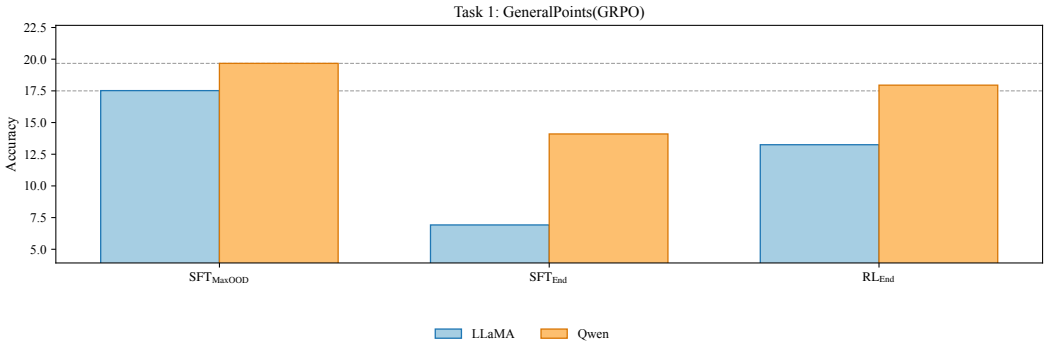


Figure 13: Results of GRPO on GeneralPoints (OOD)

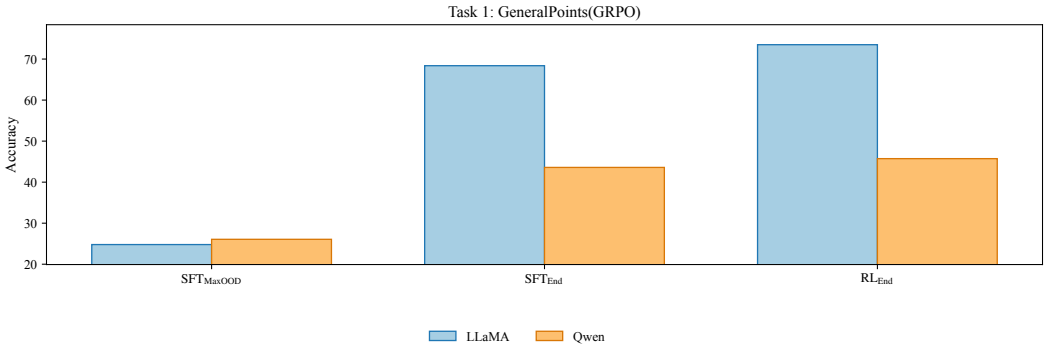


Figure 14: Results of GRPO on GeneralPoints (ID)

F.3 More Results of In-Distribution Generalization Performance

The in-distribution performance on *GeneralPoints, Navigation* and *Rank-Determinant Computation* are shown in Figure 15.

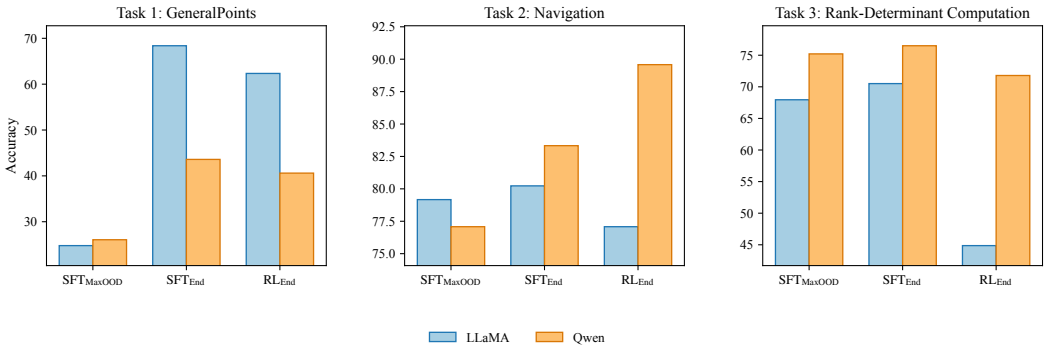


Figure 15: ID performance on three tasks

F.4 OOD Results on Other Benchmark Tasks

Table 15: OOD results at an early source-task reference, SFT_{End}, and RL_{End} on five additional benchmarks for LLaMA and Qwen. The early checkpoint is selected on the source trajectory and therefore serves as a common reference rather than a per-benchmark oracle.

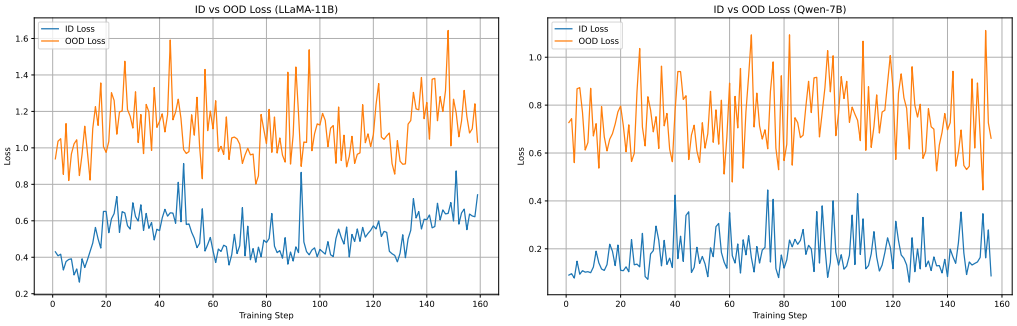
Model/Tasks		Checkpoints		
		Early reference	SFT_{End}	RL_{End}
LLaMA	ARC-Challenge	80.00	70.00	80.00
	CommonsenseQA	81.00	62.00	69.00
	GPQA	44.00	32.00	44.00
	IFEval-loose	27.00	27.00	27.00
	MMLU-Pro	46.00	36.00	37.00
Qwen	ARC-Challenge	92.00	88.00	88.00
	CommonsenseQA	79.00	73.00	75.00
	GPQA	44.00	42.00	44.00
	IFEval-loose	63.00	67.00	70.00
	MMLU-Pro	73.00	66.00	66.00

We evaluate transfer on ARC-Challenge (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), GPQA (Rein et al., 2024), IFEval-loose (Zhou et al., 2023), and MMLU-Pro (Wang et al., 2024b). Table 15 exposes benchmark-level heterogeneity. Six of the ten model–benchmark pairs show an SFT decline followed by positive RL recovery; Qwen ARC-Challenge and MMLU-Pro retain the SFT-end score after RL; LLaMA IFEval stays flat across all three points; and Qwen IFEval rises from 63 to 67 to 70. The aggregate and dense-trajectory claims therefore describe the shared checkpoint structure while preserving these task-level outcomes.

F.5 Loss of Single-Stage RL Fine-tuning

We also evaluate RL fine-tuning from the Base model as a cold start in the controlled GeneralPoints setting.

Figure 16 shows unstable Base-initialized RL loss on controlled GeneralPoints. The Base policy rarely follows the task format, producing sparse verifier success and unstable updates; SFT supplies task competence and verifier-valid candidates. The public Qwen cold-start result in Table 10 complements this loss trace: math ID scores rise, while GPQA, IFEval, and especially Safety fall, revealing a sharp ID–OOD trade-off.



(a) Loss in LLaMA (b) Loss in Qwen

Figure 16: Loss of single-stage RL fine-tuning

G Additional SVD Analysis

G.1 Direction-Sensitive Recovery and Rank-Matched Controls

The magnitude-only principal-angle analysis establishes that fine-tuning rotates dominant subspaces. We now test the direction of the RL update using three Base-relative views: signed activation projectors, an all-layer attention spectrum, and functional distances. These diagnostics use Qwen-2.5-7B/GRPO/OpenR1-Math and six OOD benchmarks.

Signed activation-projector reversal. Let $\boldsymbol{Q}_{M,k}$ contain the leading k left singular vectors of the mean-pooled activation matrix for model $M \in \{\mathrm{Base}, \mathrm{SFT}, \mathrm{RL}\}$, and let $\boldsymbol{P}_{M,k} = \boldsymbol{Q}_{M,k}\boldsymbol{Q}_{M,k}^{\top}$. We measure whether the RL projector update points from SFT toward Base with

$$c_{\mathrm{rev}}(k) = \frac{\langle \boldsymbol{P}_{\mathrm{RL},k} - \boldsymbol{P}_{\mathrm{SFT},k}, \boldsymbol{P}_{\mathrm{Base},k} - \boldsymbol{P}_{\mathrm{SFT},k} \rangle_F}{\|\boldsymbol{P}_{\mathrm{RL},k} - \boldsymbol{P}_{\mathrm{SFT},k}\|_F \, \|\boldsymbol{P}_{\mathrm{Base},k} - \boldsymbol{P}_{\mathrm{SFT},k}\|_F}. \tag{3}$$

Positive values indicate Base-directed correction. Table 16 shows positive mean correction at every rank and positive correction on all six benchmarks at top-16 and top-32.

OOD benchmark	Top-4	Top-8	Top-16	Top-32
GPQA	+0.117	+0.169	+0.084	+0.113
IFEval	−0.195	+0.048	+0.027	+0.054
MMLU-Pro	+0.275	−0.046	+0.034	+0.012
SuperGPQA	+0.086	−0.032	+0.064	+0.106
Safety	+0.193	+0.113	+0.052	+0.133
TruthfulQA-MC	+0.189	+0.189	+0.250	+0.202
Mean	+0.111	+0.074	+0.085	+0.103

Table 16: Signed reversal cosine of the RL activation-projector update toward the Base projector.

All-layer key-projection spectrum. For each model, concatenate the singular values of its attention key-projection matrices across all $L = 28$ Transformer layers, $\boldsymbol{s}_M = \mathrm{concat}_{\ell=1}^{L} \sigma(\boldsymbol{W}_{K,M}^{(\ell)})$. The spectral correction cosine is

$$c_K = \frac{\langle \boldsymbol{s}_{\mathrm{RL}} - \boldsymbol{s}_{\mathrm{SFT}}, \boldsymbol{s}_{\mathrm{Base}} - \boldsymbol{s}_{\mathrm{SFT}} \rangle}{\|\boldsymbol{s}_{\mathrm{RL}} - \boldsymbol{s}_{\mathrm{SFT}}\|_2 \|\boldsymbol{s}_{\mathrm{Base}} - \boldsymbol{s}_{\mathrm{SFT}}\|_2} = 0.146. \tag{4}$$

Three thousand layer-block resamples give a 95% interval of $[0.114, 0.177]$. The independent parameter-spectrum diagnostic therefore agrees with the signed activation result.

Behavioral and functional contraction. Table 17 aligns the positive OOD pass@1 gain with two Base-relative function-space changes. Hidden-state contraction is the percentage decrease

in mean-pooled activation distance from Base, and KL contraction is the percentage decrease in $D_{\mathrm{KL}}(p_{\mathrm{Base}} \| p_M)$ for the next-token distribution. All six benchmarks move closer to Base under both measures while their OOD pass@1 increases.

OOD benchmark	Pass@1 gain (pp)	Hidden-state contraction [95% CI]	KL contraction [95% CI]
GPQA	+1.120	13.940% [12.720, 15.230]	10.660% [2.370, 18.730]
IFEval	+0.180	4.540% [3.940, 5.170]	8.240% [4.040, 12.680]
MMLU-Pro	+5.720	8.580% [7.610, 9.560]	9.190% [3.380, 14.790]
SuperGPQA	+5.680	11.170% [9.810, 12.610]	11.710% [1.500, 20.010]
Safety	+2.800	4.890% [4.470, 5.320]	8.610% [3.340, 13.810]
TruthfulQA-MC	+5.700	2.230% [1.820, 2.630]	7.280% [0.580, 13.790]

Table 17: OOD improvement and functional contraction toward the Base model for Qwen-2.5-7B after GRPO. Bracketed intervals use evaluation-example resampling.

Partial-rank intervention. We probe direction identity with a hybrid matrix that retains every SFT coefficient, substitutes the leading k Base direction pairs, and leaves the SFT tail unchanged:

$$\boldsymbol{W}^{UV}_{\mathrm{recover}}(k) = \sum_{i=1}^{k} \sigma_i^S \boldsymbol{u}_i^B (\boldsymbol{v}_i^B)^\top + \sum_{i=k+1}^{r} \sigma_i^S \boldsymbol{u}_i^S (\boldsymbol{v}_i^S)^\top. \tag{5}$$

Here B and S denote Base and SFT, and r is compact rank. The σ_i^S values are retained expansion coefficients; mixed Base/SFT direction sets form a hybrid expansion. A rank-matched random control replaces the same number of direction pairs and averages five independent random-rank seeds. For the final table column, if B is Base OOD accuracy and R is post-intervention OOD accuracy, forgetting relative to Base is $100(B - R)/B$.

Model / shift	k	Directions	Recovered OOD (%)	Top-k Fisher block (%)	Top-k Fisher energy (%)	Forgetting vs. Base (%)
Llama-3.2-11B / rule-changed cards	64	Top	7.260	0.024	0.467	57.294
	64	Random mean	7.816	–	–	54.024
	256	Top	7.360	0.391	3.141	56.706
	256	Random mean	8.600	–	–	49.412
	512	Top	8.970	1.563	8.061	47.235
	512	Random mean	7.144	–	–	57.976
	768	Top	11.970	3.516	13.458	29.588
	768	Random mean	7.480	–	–	56.000
Qwen2.5-7B / Math →instruction	64	Top	38.462	0.032	0.978	31.100
	64	Random mean	30.000	–	–	46.259
	256	Top	42.308	0.510	5.932	24.210
	256	Random mean	32.692	–	–	41.436
	512	Top	40.385	2.041	13.265	27.655
	512	Random mean	27.692	–	–	50.393
	768	Top	44.231	4.592	21.300	20.766
	768	Random mean	28.846	–	–	48.326

Table 18: Rank-matched direction intervention. Llama Base/SFT OOD accuracies are 17.000/6.920; Qwen Base/SFT accuracies are 55.823/41.590. "Top" uses Equation 5; "Random mean" averages five independent direction-set seeds at the same rank. Fisher block is the selected fraction of SVD product-coordinate space; Fisher energy is the empirical-Fisher energy captured by that block.

At fixed rank, pretrained directions outperform the random-control mean at every Qwen rank and at the two larger Llama ranks. Fisher energy rises with k, and the largest intervention yields 11.970% Llama OOD accuracy and 44.231% Qwen OOD accuracy.

Natural-scale value–direction factorial. The rank-matched controls above ask whether pretrained direction identity matters for a direction intervention. We additionally decompose the naturally realized Base-to-SFT spectral change without rescaling either component. At the Qwen2.5-7B/OpenR1 SFT checkpoint, outcome-independent stability criteria select spectral blocks from 168 attention and MLP matrices. For compact notation, the first letter indexes singular values and the second indexes

singular-vector directions, with A denoting the Base component and B the SFT component:

	singular values	singular-vector directions
BB	$\boldsymbol{\Sigma}_{\mathrm{SFT}}$	$\boldsymbol{D}_{\mathrm{SFT}}$
AB	$\boldsymbol{\Sigma}_{\mathrm{Base}}$	$\boldsymbol{D}_{\mathrm{SFT}}$
BA	$\boldsymbol{\Sigma}_{\mathrm{SFT}}$	$\boldsymbol{D}_{\mathrm{Base}}$
AA	$\boldsymbol{\Sigma}_{\mathrm{Base}}$	$\boldsymbol{D}_{\mathrm{Base}}$

Here $\boldsymbol{D}$ denotes the paired left and right singular-vector directions. All parameters outside the selected blocks remain at the SFT checkpoint, so AA is a hybrid and is not the Base model. For an accuracy functional f, the symmetric factorial attributions and interaction are

$$\phi_{\Sigma} = \frac{1}{2}[(f_{AB} - f_{BB}) + (f_{AA} - f_{BA})], \qquad \phi_{\mathrm{dir}} = \frac{1}{2}[(f_{BA} - f_{BB}) + (f_{AA} - f_{AB})],$$
$$I = f_{AA} - f_{AB} - f_{BA} + f_{BB}.$$

	Accuracy (%)				Effect (pp) [95% CI]			
Benchmark	BB	AB	BA	AA	ϕ_{Σ}	ϕ_{dir}	$BA - AB$	I
GPQA	31.62	32.44	33.48	34.38	+0.86 [−1.04, 2.75]	+1.90 [−0.60, 4.35]	+1.04 [−2.16, 4.09]	+0.07 [−4.02, 4.24]
IFEval	55.39	55.08	63.89	62.85	−0.68 [−1.97, 0.55]	+8.13 [5.70, 10.60]	+8.81 [6.10, 11.58]	−0.74 [−3.27, 1.91]
MMLU-Pro	66.19	60.48	62.86	71.43	+1.43 [−2.86, 5.71]	+3.81 [−1.90, 9.76]	+2.38 [−4.29, 9.05]	+14.29 [6.67, 22.38]
SuperGPQA	28.15	26.91	26.83	28.07	+0.00 [−1.89, 1.93]	−0.08 [−2.51, 2.30]	−0.08 [−3.13, 3.05]	+2.47 [−1.48, 6.50]
Safety	70.60	69.90	71.87	72.10	−0.23 [−0.99, 0.51]	+1.73 [−0.47, 3.85]	+1.97 [−0.40, 4.28]	+0.93 [−0.47, 2.37]
TruthfulQA-MC	61.16	61.74	62.87	62.72	+0.22 [−1.00, 1.46]	+1.34 [−0.49, 3.14]	+1.12 [−1.07, 3.31]	−0.73 [−3.07, 1.66]
Equal-six OOD macro	**52.19**	**51.09**	**53.63**	**55.26**	**+0.27** [−0.63, 1.19]	**+2.81** [1.56, 4.09]	**+2.54** [1.07, 4.04]	**+2.72** [1.01, 4.49]

Table 19: Natural-scale factorial on selected stable spectral blocks of 168 Qwen2.5-7B matrices. The first four columns report accuracy (%); the remaining columns report percentage-point effects with paired 95% confidence intervals. Here, ϕ_{Σ} and ϕ_{dir} are the symmetric factorial attributions to singular values and singular-vector directions, respectively; $BA - AB = \phi_{\mathrm{dir}} - \phi_{\Sigma}$ is the direction-minus-value contrast, and I is the value–direction interaction. Accuracy is averaged over decoder streams, and intervals use paired prompt-cluster resampling after averaging streams. The OOD macro gives equal weight to the six listed benchmarks. All parameters outside the selected spectral blocks remain at the SFT checkpoint, so AA is a hybrid and is not the Base checkpoint.

On the equal-six OOD macro, restoring Base directions alone raises accuracy from 52.19 (BB) to 53.63 (BA), whereas restoring Base singular values alone lowers it to 51.09 (AB). The symmetric attribution is $+2.81$ pp (95% CI $[1.56, 4.09]$) for directions and $+0.27$ pp (95% CI $[-0.63, 1.19]$) for values; the direction-minus-value contrast is $+2.54$ pp (95% CI $[1.07, 4.04]$). On IFEval, restoring directions raises accuracy from 55.39 to 63.89, while restoring values alone gives 55.08. The IFEval $BA - AB$ gap is positive under all three decoder streams ($+9.98$, $+7.58$, and $+8.87$ pp). The positive OOD-macro interaction ($+2.72$ pp, 95% CI $[1.01, 4.49]$) indicates that compatible coefficients and directions can also contribute jointly; the direction marginal nevertheless substantially exceeds the value marginal.

As an ID sign check, AIME-25 moves from 12.22 (BB) to 10.00 (BA) and 6.67 (AA), while AB remains at 12.22. This opposite movement is consistent with partially reversing SFT specialization, although its intervals include zero and we do not use it as evidence for a statistically resolved ID cost. Together with the geometry-matched random-direction comparison below, the factorial separates two claims: at the magnitudes realized by SFT, direction changes carry more of the OOD transition than value changes alone, and the identity of the restored pretrained directions matters beyond perturbation magnitude.

Geometry-matched Base-direction dose response. We further test direction identity from Qwen2.5-7B OpenR1 SFT checkpoint 19,200. Across 168 attention and MLP matrices, we apply five doses of a Base-directed SVD substitution with target rank 768, bounded by each matrix's compact rank. Each dose is paired with three random orthogonal controls matched by layer, matrix family, effective rank, realized BF16 Frobenius displacement, and the complete left- and right-space principal-angle spectra. The unmodified checkpoint attains 55.268% IFEval accuracy on all 541 prompts; this is the checkpoint-specific baseline for the dose-controlled comparison.

Dose	SVD norm	Random norm	SVD Δ (pp)	Random Δ (pp)	SVD$-$random (pp)	Paired 95% CI
0.125	4.588	4.588	+2.773	$+0.185 \pm 0.370$	+2.588	$[-0.185, 5.360]$
0.250	9.028	9.046	+4.067	$+0.370 \pm 0.847$	+3.697	[0.678, 6.778]
0.500	17.756	17.758	+5.915	$+0.678 \pm 0.282$	+5.237	[2.341, 8.195]
0.750	26.528	26.529	+8.688	$+0.801 \pm 1.337$	+7.887	[4.868, 10.906]
1.000	35.317	35.318	+10.536	$+0.062 \pm 0.949$	+10.474	[7.209, 13.740]

Table 20: Geometry-matched Qwen2.5-7B IFEval dose response from OpenR1 SFT checkpoint 19,200. Norms are realized BF16 Frobenius displacements and match within 0.2% in the reported aggregate. Changes are accuracy differences from the 55.268% SFT baseline on 541 prompts. Random changes are means $\pm$ one SD across three matched orthogonal controls; the final column reports the paired 95% confidence interval for the Base-directed change minus the random-control mean.

The Base-directed gain increases monotonically with intervention norm from $+2.773$ to $+10.536$ pp, whereas the random-control mean remains between $+0.062$ and $+0.801$ pp. The Base-directed advantage rises from $+2.588$ to $+10.474$ pp, and the paired intervals exclude zero at doses 0.25 through 1.00. Near-repeated singular values are grouped as invariant subspaces using 10^{-4} relative-gap clustering, making the comparison stable under basis rotations within clustered blocks and at the rank boundary.

In the separate soft subspace-regularized layer-7 W_K comparison from Table 12, the selected tail coordinate is 13.929° for Base$\rightarrow$SFT and 1.025° for the subsequent SFT$\rightarrow$RL stage. The signed activation and all-layer key-spectrum results establish the Base-directed component, while the rank-matched and dose-matched interventions connect pretrained direction identity to functional OOD recovery.

G.2 SVD Notation

For a parameter matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, we write its singular-value decomposition as $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ contain the left and right singular vectors and $\boldsymbol{\Sigma}$ contains non-negative singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. In our analysis, the singular values measure the strength of each mode and the singular vectors define its directions. We apply this decomposition to self-attention and MLP matrices in LLaMA and Qwen.

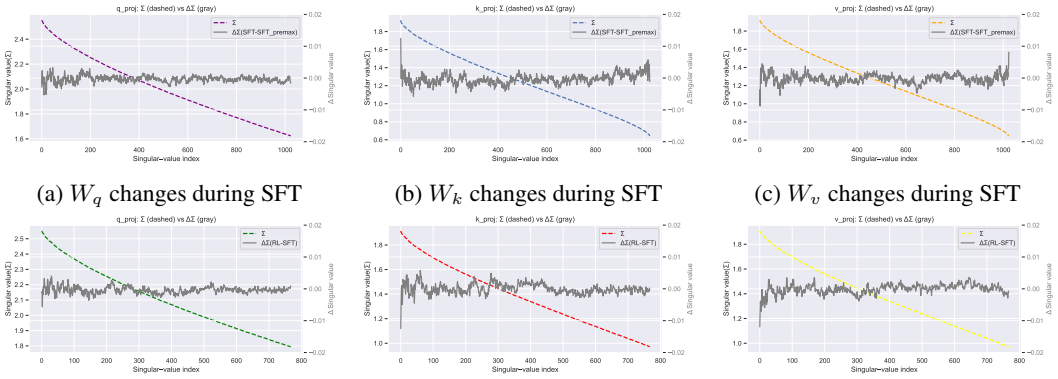
G.3 Changes of Singular Values

To measure how SFT and RL reshape parameter spectra, we analyze the singular values of $\boldsymbol{W_q}, \boldsymbol{W_k}, \boldsymbol{W_v}$ and their stage-wise differences ($\Delta\sigma_i = \sigma_i^{\text{SFT}_{\text{End}}} - \sigma_i^{\text{SFT}_{\text{MAXOOD}}}$). Figure 17 shows a representative LLaMA layer-5 measurement in which the $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ singular-value differences after both stages span 0–0.005 as a low-magnitude, approximately zero-centered signal. The broader value-restoration diagnostics in Figures 5 and 6 measure the corresponding functional effect across layers and ranks.

G.4 Exploring the Rotation of Singular Vector with Principal Angles

There are two ways to measure singular-vector changes during fine-tuning: vector-level metrics and subspace-level metrics.

Principal angles (or canonical angles) quantify how far two subspaces are within the same Euclidean space. To quantify the differences between the subspaces spanned by the singular vectors of base model $\boldsymbol{W}_{\text{Base}}$ and fine-tuned model $\boldsymbol{W}_{\text{FT}}$, we measure the amount of rotations between two subspaces by how much their dominant singular vector directions have *rotated*, which is a commonly used method in machine learning (Huang et al., 2015; Vahidian et al., 2023) and numerical computation (Björck & Golub, 1973). We provide a brief introduction and we take the left singular vectors for example and the computation includes,



(a) W_q changes during SFT (b) W_k changes during SFT (c) W_v changes during SFT

(d) W_q changes during RL (e) W_k changes during RL (f) W_v changes during RL

Figure 17: Singular-value changes in the `q_proj`, `k_proj`, and `v_proj` matrices of self-attention layer 5 (zero-indexed; `layers[5].self_attn`) in `LLaMA-3.2-11B-Vision`. Panels (a)–(c) show SFT-stage differences for W_q, W_k, and W_v, while panels (d)–(f) show the corresponding RL-stage differences. SFT is measured from $\text{SFT}_{\text{MaxOOD}}$ to SFT_{End}; RL begins at SFT_{End}.

(i) SVD. For each matrix, we compute the full SVD and retain the leading k blocks for each reported principal-angle rank,

$$W = U \Sigma V^\top, \qquad U \in \mathbb{R}^{m\times r},\ V \in \mathbb{R}^{n\times r}, \tag{2.4.1}$$

where the columns of U and V are orthonormal and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$, r is the rank.

(ii) Computation of Principal Angles Between Subspaces (PABS). Let $\boldsymbol{U}_{\text{Base}}, \boldsymbol{U}_{\text{FT}} \in \mathbb{R}^{m\times k}$ be the left singular blocks from the previous step. Define $\boldsymbol{M} := \boldsymbol{U}_{\text{Base}}^\top \boldsymbol{U}_{\text{FT}} \in \mathbb{R}^{k\times k}$. Since both blocks have orthonormal columns, the singular values of M lie in $[0, 1]$ (Björck & Golub, 1973). Suppose the SVD of $\boldsymbol{M}$ is

$$\boldsymbol{M} = \boldsymbol{U}_M \,\text{diag}(s_1, \ldots, s_k)\, \boldsymbol{V}_M^\top,$$

the *principal angles* $\theta_i \in [0, \pi/2]$ between the column spaces of $\boldsymbol{U}_{\text{Base}}$ and $\boldsymbol{U}_{\text{FT}}$ are

$$\theta_i \ = \ \arccos(s_i), \quad i = 1, \ldots, k. \tag{2.4.2}$$

The full decomposition costs $O(\min\{m, n\}^3)$; the $k \times k$ overlap SVD costs $O(k^3)$. An identical procedure on $\boldsymbol{V}_{\text{Base}}, \boldsymbol{V}_{\text{FT}}$ yields angles for the right subspaces. In practice we clamp the numerical values of s_i to $[0, 1]$ before calling $\arccos$ to avoid floating-point overflow. Principal angles measure the tilt between singular-vector subspaces: $\theta_i = 0$ means that the corresponding direction is preserved, whereas values approaching $\pi/2$ indicate maximal misalignment.

Advantages of PABS

- **Numerical Stability:** When two nearby singular values swap order after a small update, pairwise vector cosine can show a large angle despite a small matrix change. PABS remains stable under this reordering.
- Cosine similarity between singular vectors only compares one dimension at a time, without accounting for interdependence between directions. PABS derives angles that reflect the relative orientation of the entire subspace, providing a more informative measure than isolated vector-to-vector comparisons.
- PABS provides a true metric for comparing subspaces and measures alignment or divergence holistically.

We use principal angles to measure subspace rotation during SFT and RL. For Figure 18, we plot every ordered right-space principal angle for the single layer-0 k_{proj} matrix from Base to SFT_{End} and from Base to RL_{End}. Both per-angle spectra increase smoothly from roughly 25–30 degrees in the tail, establishing that both stages rotate this parameter subspace. Table 12

reports selected angle coordinates for a separate fixed-top-512 layer-7 W_K comparison. Neither result averages across layers or matrix families; the all-layer aggregate is the key-spectrum correction in Equation 4. Magnitude alone merges distinct directions at many ranks. The signed projector and key-spectrum tests in Appendix G.1 resolve that direction: RL points from SFT toward Base across activation and parameter views. Rank-matched interventions then connect the identity of those pretrained directions to recovered OOD accuracy.

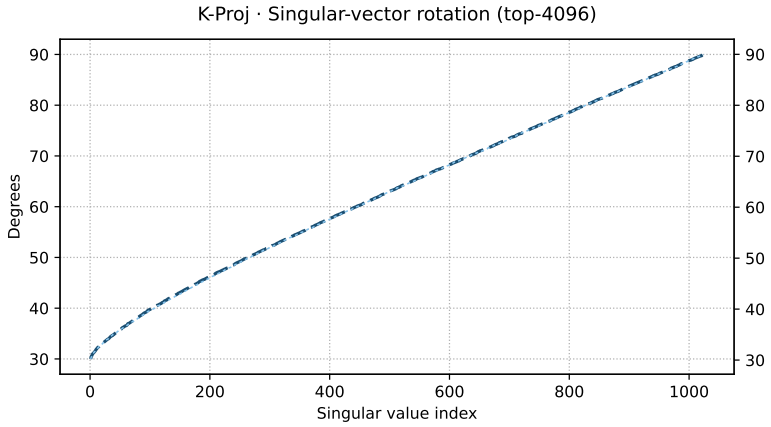
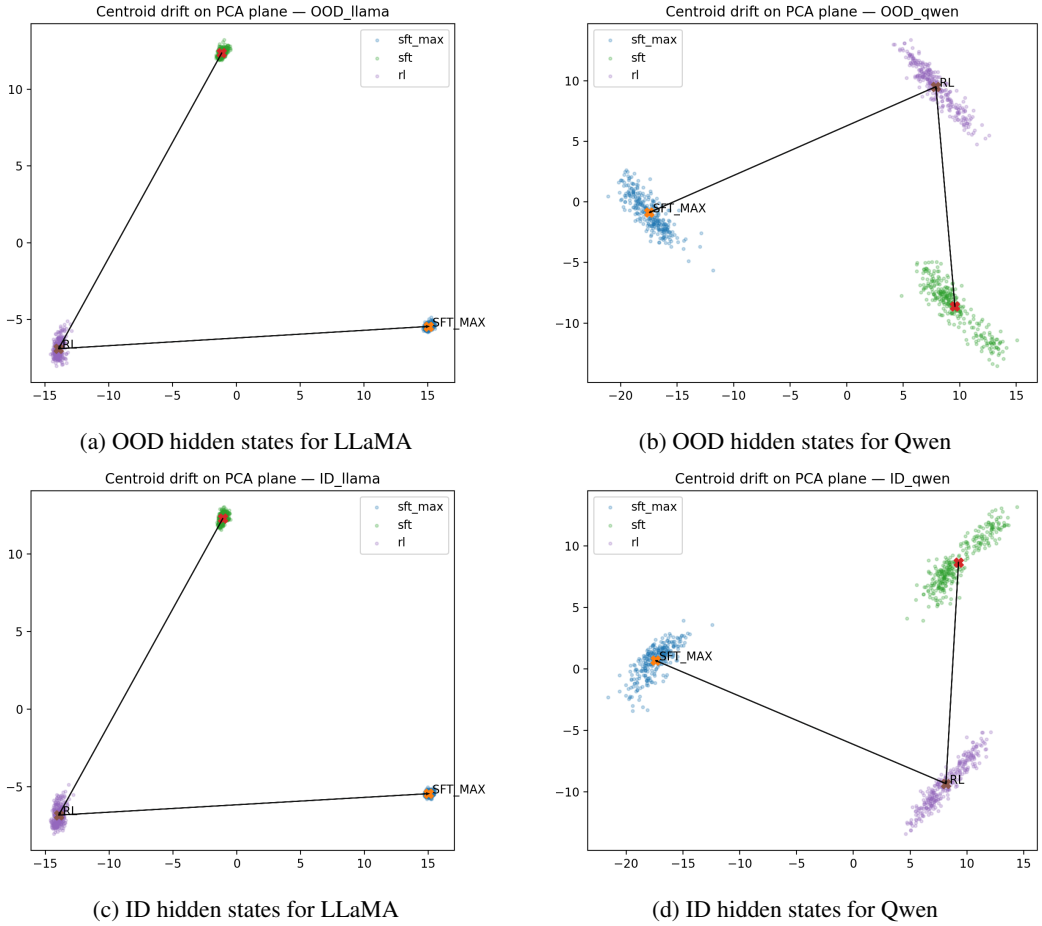


Figure 18: Individual right-space principal-angle spectra for the layer-0 k_{proj} matrix from Base to $\mathrm{SFT}_{\mathrm{End}}$ and from Base to $\mathrm{RL}_{\mathrm{End}}$. Each point is one ordered canonical angle for this single matrix; Table 12 reports selected coordinates from the separate layer-7 mitigation comparison.

G.5 PCA Visualization of Embedding Shifts



(a) OOD hidden states for LLaMA

(b) OOD hidden states for Qwen

(c) ID hidden states for LLaMA

(d) ID hidden states for Qwen

Figure 19: PCA visualization of the hidden representations at checkpoints $\text{SFT}_{\text{MaxOOD}}$, SFT_{END} and RL_{END}.

We activate hidden states with 300 ID and 300 OOD prompts at each checkpoint, fit PCA to the representation matrix, and visualize the first two principal components. In this two-dimensional projection, the RL_{End} centroid lies farther from $\text{SFT}_{\text{MaxOOD}}$ than the SFT_{End} centroid does. This descriptive view uses $\text{SFT}_{\text{MaxOOD}}$ as its reference and retains only two components. The direction-sensitive tests in Appendix G.1 instead use the pretrained Base as the reference and measure full-dimensional projector, hidden-state, distributional, and parameter-spectrum contraction. The two analyses therefore describe different geometric relations.

G.6 Early Verifier-Error Decomposition

Setting We evaluate 234 OOD examples at each early SFT checkpoint. Each trajectory permits up to six attempts under the fixed verifier protocol. OOD accuracy is the percentage of trajectories with at least one verifier-correct final answer.

Verifier and check order (per step) 1. Format parse: if parse fails → ILLEGAL_FORMAT. Stop other checks for that step.
2. Number check: if numbers in the formula fall outside the supplied set or use the wrong count → INCORRECT_NUMBER.
3. Solution check: if no valid "final answer" after format checking → NO_SOLUTION.
4. Aggregation: if ≥ 2 of the above are true for a step → also count AGGREGATED_ERR.

How we compute metrics - CORRECT_SOLUTION(CS): a trajectory is correct if any of its six attempts ends with a verifier-correct final answer; accuracy is the percentage over 234 trajectories.
- Step-level rates (NO_SOLUTION(NS), ILLEGAL_FORMAT(IF), INCORRECT_NUMBER(IN), AGGREGATED_ERR(AE)): count steps with the label divided by all steps. Rates can co-occur, so their total can exceed 100%.

ck	CS	NS	IF	IN	AE
10	0.00	18.95	80.06	0.78	0.21
20	1.28	33.57	54.62	7.59	4.01
30	10.26	63.68	0.99	20.59	12.92
40	9.40	84.07	0.23	7.13	6.90
50	12.39	91.09	0.08	4.30	2.30
60	16.24	90.95	0.08	3.70	2.28
70	16.24	86.44	0.00	5.34	5.27
80	13.25	90.80	0.00	4.49	2.32
90	14.10	82.64	0.00	7.97	6.80
100	17.09	87.76	0.00	4.74	4.34
110	15.38	88.54	0.00	5.69	2.92
120	15.38	91.82	0.00	4.17	1.18
130	15.81	90.66	0.00	3.14	3.30
140	17.52	90.13	0.00	3.82	2.87
150	14.96	92.55	0.00	2.90	1.80

Table 21: OOD trajectory accuracy and step-level verifier-error rates over 234 evaluation examples during early SFT. Each trajectory permits up to six attempts.

Observation - CORRECT (OOD) grows from 0.00% (ck=10) to $\approx 17.00\%$ (ck=100–150).
- ILLEGAL_FORMAT drops to 0.00% by ck$\geq$70 (better syntax).
- INCORRECT_NUMBER falls after peaking near ck=30 (better number use).
- AGGREGATED_ERR stays low and trends down.

Relation to the loss–accuracy discrepancy - Different targets: CE fits ID train tokens; CORRECT measures end-to-end OOD success under verifier scoring and multiple attempts.
- Structure over tokens: cleaner format and number use can boost pass@6 even if CE rises.
- Search effect: as more attempts are valid, at-least-one-success increases.
- Scope: checkpoints 10–150 capture early capability acquisition through the OOD peak; the later decline is measured by the dense trajectories in Figures 2c and 3a.

G.7 Advantage Distribution Metrics

For each SFT checkpoint k, we collect the advantage samples produced during RL rollouts and compute basic statistics: center μ_k, standard deviation σ_k, and skewness (Zwillinger & Kokoska, 1999). To capture distribution shape beyond the first moments, we also use entropy and KL divergence to a matched normal distribution.

Entropy We use entropy to measure flatness and uniformity. Flat advantage distributions without pronounced modes have higher entropy than sharply concentrated distributions (Petty, 2018). For a discrete approximation $\hat{p}_k$ over bins b, we compute

$$H(\hat{p}_k) = -\sum_b \hat{p}_k(b) \log \hat{p}_k(b).$$

For finite discrete support, the uniform distribution has maximum entropy (Hyvärinen, 1997).

KL divergence to a matched normal distribution We use KL divergence to measure non-Gaussian structure (Hyvärinen, 1997; Hyvärinen & Oja, 2000), or negentropy (Hyvärinen, 2013). For checkpoint k, the matched normal distribution is $q_k = N(\mu_k, \sigma_k^2)$, where μ_k and σ_k are estimated from the advantage samples. For the same discrete approximation, we compute

$$D_{\mathrm{KL}}(\hat{p}_k \| q_k) = \sum_b \hat{p}_k(b) \log \frac{\hat{p}_k(b)}{q_k(b)}.$$

Given fixed second moment or variance, a Gaussian maximizes differential entropy (Tse, 2017), so larger KL indicates more non-Gaussian structure, such as sharper peaks, multiple modes, heavier tails, or asymmetry (Freeman & Dale, 2013; Silverman, 1981).

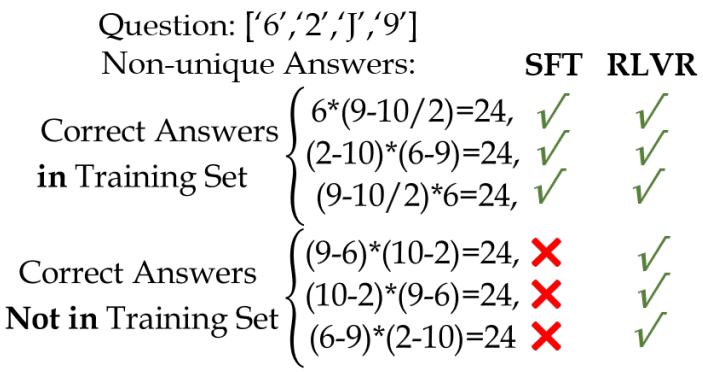
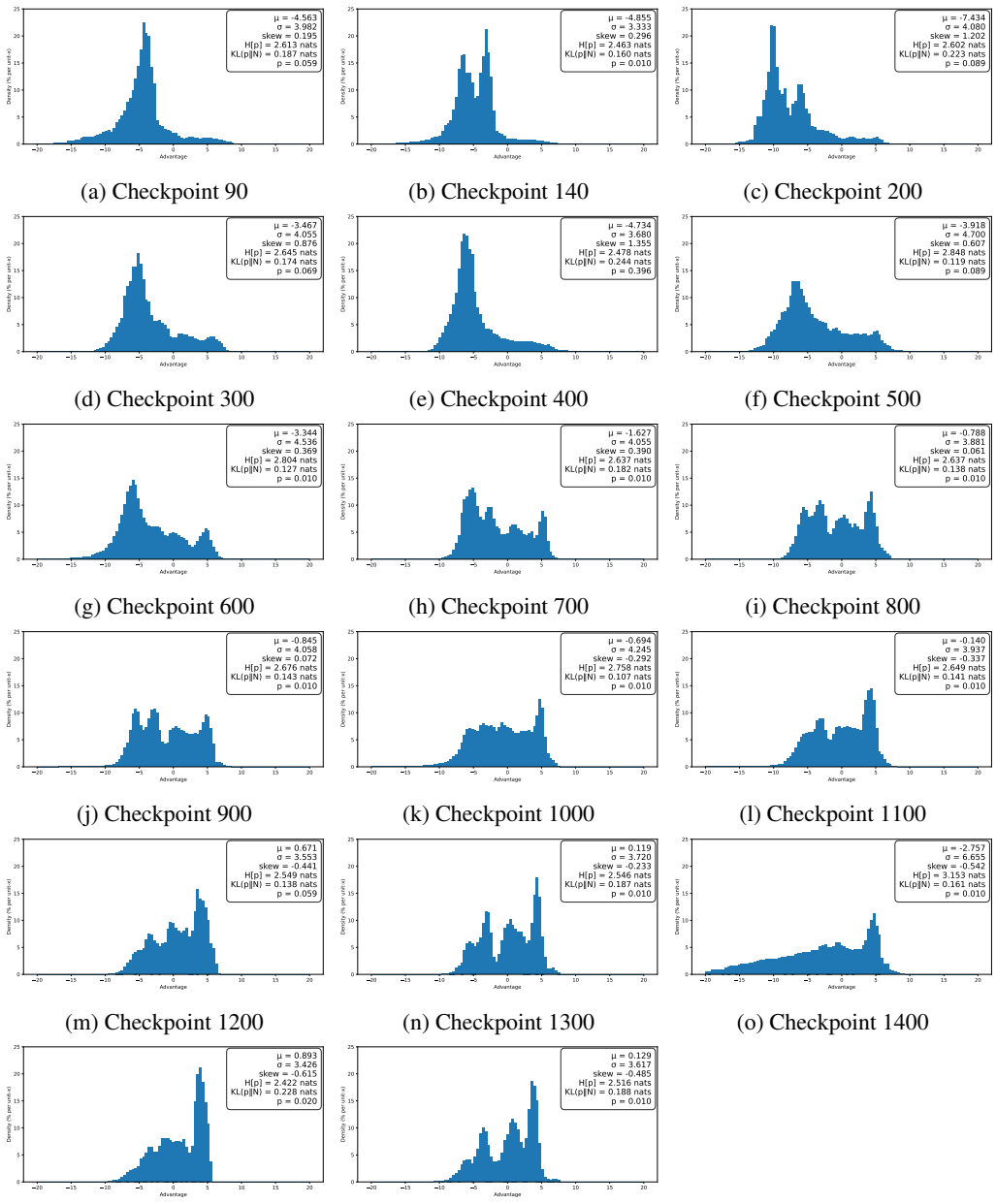


Figure 20: Verifier-correct GeneralPoints derivations for the same input. Outcome-level RL can reinforce either solution, whereas token-level SFT concentrates supervision on the demonstrated sequence.

G.8 Full Results of Advantage Distribution



(a) Checkpoint 90 (b) Checkpoint 140 (c) Checkpoint 200

(d) Checkpoint 300 (e) Checkpoint 400 (f) Checkpoint 500

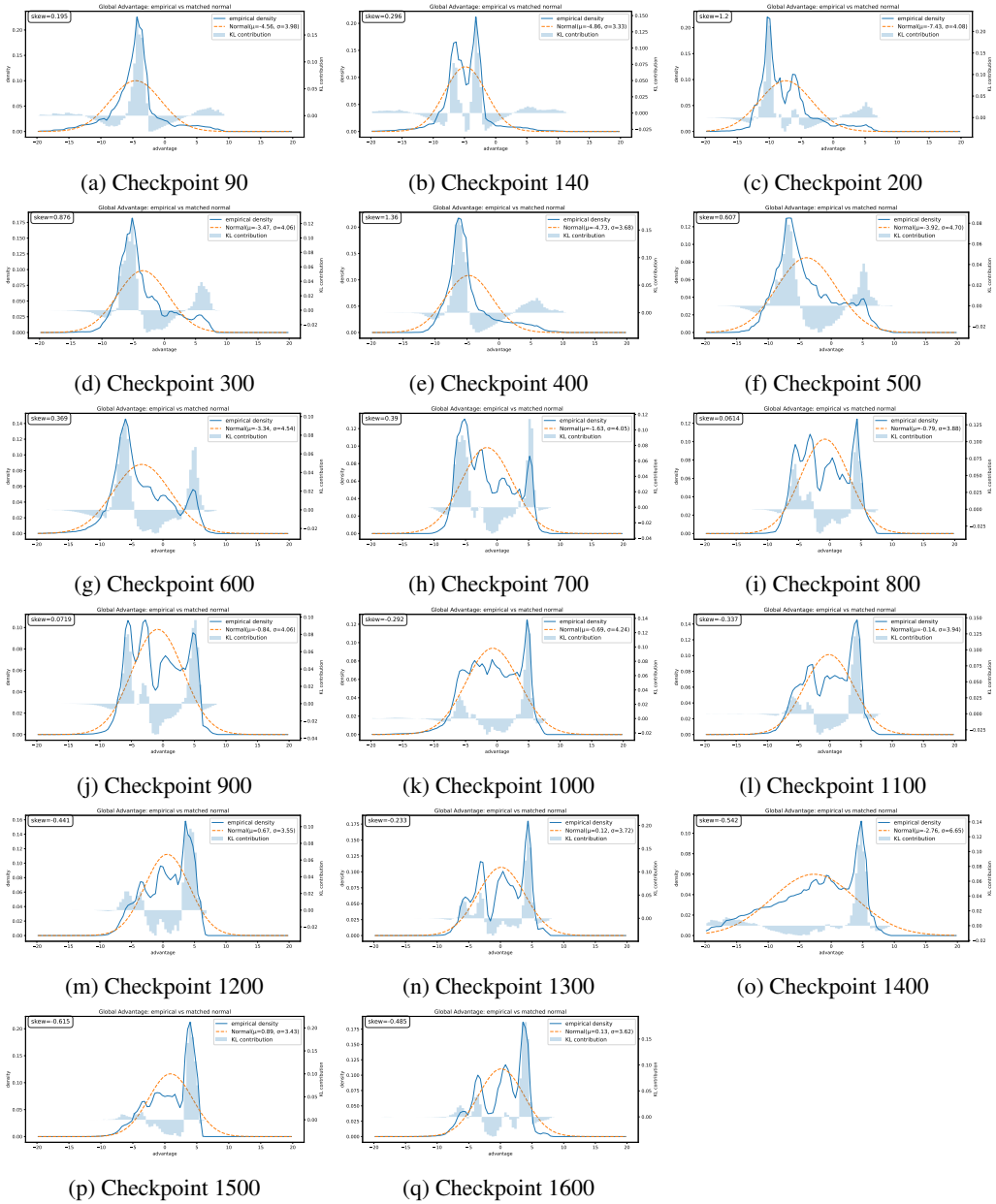
(g) Checkpoint 600 (h) Checkpoint 700 (i) Checkpoint 800

(j) Checkpoint 900 (k) Checkpoint 1000 (l) Checkpoint 1100

(m) Checkpoint 1200 (n) Checkpoint 1300 (o) Checkpoint 1400

(p) Checkpoint 1500 (q) Checkpoint 1600

Figure 21: Full checkpoint-wise advantage histograms. Each panel reports the sampled distribution and its skewness for one LLaMA/PPO initializer.



(a) Checkpoint 90 (b) Checkpoint 140 (c) Checkpoint 200

(d) Checkpoint 300 (e) Checkpoint 400 (f) Checkpoint 500

(g) Checkpoint 600 (h) Checkpoint 700 (i) Checkpoint 800

(j) Checkpoint 900 (k) Checkpoint 1000 (l) Checkpoint 1100

(m) Checkpoint 1200 (n) Checkpoint 1300 (o) Checkpoint 1400

(p) Checkpoint 1500 (q) Checkpoint 1600

Figure 22: Full checkpoint-wise advantage densities compared with their matched normal distributions. Each panel reports the corresponding KL divergence.