# DOES SCALING REINFORCEMENT LEARNING REALLY REQUIRE MORE TRAINING?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Scaling reasoning typically spends more compute on reinforcement learning (RL) or on inference. We show that a completed RL training history can yield policies stronger than the checkpoints visited by its optimizer. We call this *policy-space scaling*: expanding the deployable policy set accessible from a fixed RL history, without extending training or increasing per-query inference computation. We instantiate it with SURGE (Scaling Up RL Gradient-free via Eigenspace fusion). SURGE combines two checkpoints from the same RL run: a high-accuracy anchor and a competitive donor that generates shorter responses. It expresses both checkpoints as changes from their shared initialization, then spectrally decomposes the anchor's update to retain its dominant component and incorporate the donor's complementary component. With a fixed target for how much of the anchor update to retain, SURGE determines the block size from the weights without testing candidate policies. We evaluate two 1.5B mathematical-reasoning histories, DeepSeek and Nemotron, and one 7B coding history, OLMo. SURGE improves benchmark-average accuracy over both input checkpoints while using fewer reasoning tokens than the anchor. It reaches 54.17% on DeepSeek AIME24 against a measured native maximum of 50.83%, and 83.7% on OLMo HumanEval+ against 82.8%. These gains exceed the observed training curves. Geometric controls support the importance of RL-update structure beyond weight displacement or token reduction alone. Each constructed model runs as a single policy. Our findings identify stored RL history as a reusable scaling resource: the capability available from a training run need not end at its best checkpoint.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards has become a practical route to improving reasoning in language models (Shao et al., 2024; DeepSeek-AI, 2025). When performance is insufficient, the usual response is to continue optimization or allocate more computation to inference. Both extend a computation process. We ask whether the same training run already contains enough structure to construct a stronger policy. A training trajectory records where an optimizer has been; it need not exhaust the capability obtainable from the updates it has computed.

This distinction changes what counts as the output of RL. Checkpoint selection treats a run as a finite list of complete policies. Viewed in parameter space, the same history also contains distinct updates relative to a common initialization. Those updates can support policies that the optimizer never visited. Because reasoning performance is nonlinear in the weights, recombining update components need not average the capabilities of their sources. The relevant question is therefore whether stored history supports a higher accuracy ceiling than native checkpoint selection alone.

We study *policy-space scaling*: expanding the deployable policy set accessible from a fixed RL history. Training-time scaling expands what the optimizer can learn by extending optimization. Test-time scaling expands what a fixed policy can compute by extending inference. Policy-space scaling expands what a completed training history can deploy by constructing additional policies from already-computed updates. Offline policy construction and evaluation have a cost, but the resulting model has the same architecture and runs as one policy at deployment.

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

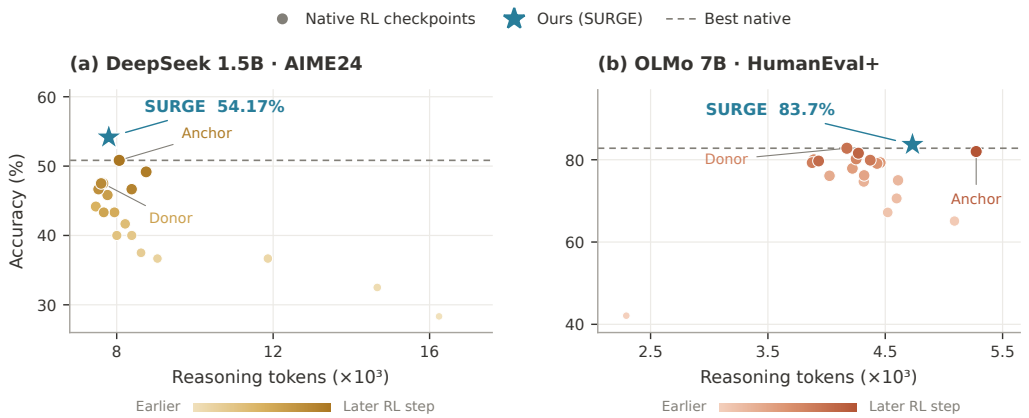



Figure 1: **Scaling beyond observed RL checkpoints, from 1.5B math to 7B code.** Circles encode native RL step (color) and accuracy (size); SURGE exceeds the best measured native accuracy (dashed line) with fewer tokens than the anchor. Each panel shows 19 checkpoints on linear axes. Token counts are in thousands.

Our contributions are threefold:

- We show that an RL history can support policies stronger than its best measured checkpoint, and formulate *policy-space scaling*: expanding the deployable policy set from already-computed updates, without further RL or a larger inference budget.
- We develop SURGE to instantiate this axis through delta-first spectral construction, preserving an anchor's protected component and admitting a donor's complementary component. Given a source pair and an empirical energy target, weight-only calibration determines the setting without an accuracy sweep.
- We demonstrate this opportunity across 1.5B math and 7B code. SURGE exceeds the best of 19 measured native checkpoints on DeepSeek AIME24 (54.17% vs. 50.83%) and OLMo HumanEval+ (83.7% vs. 82.8%; Figure 1). Across nine math benchmarks, average accuracy improves over the anchors by 2.08/1.78 points for DeepSeek/Nemotron, with fewer tokens on every benchmark. Geometric controls support the importance of RL-update structure.

## 2 METHODOLOGY

### 2.1 AN RL TRAJECTORY AS A RESOURCE FOR SCALING

Let $\theta_0$ initialize RL and $\mathcal{C}_T = \{\theta_t : t \in \mathcal{I}_T\}$ contain its stored checkpoints with nonzero RL updates through step $T$. Native checkpoint selection is restricted to $\mathcal{C}_T$. A construction $F$ acting on two compatible checkpoints opens a larger policy family,

$$\mathcal{P}_T = \mathcal{C}_T \cup \{F(\theta_a, \theta_b; f) : a, b \in \mathcal{I}_T,\ a \neq b,\ f \in \mathcal{F}\}. \tag{1}$$

Training-time scaling extends the history; policy-space scaling explores $\mathcal{P}_T$ while holding that history fixed. Equation 1 defines a candidate space, not a requirement to enumerate all pairs. We select an accuracy-strong checkpoint as the *anchor* and a competitive, shorter-response checkpoint as the *donor*, using source-policy statistics already available along the RL trajectory. These roles need not follow a temporal ordering or a per-benchmark accuracy ranking. Source selection is distinct from the weight-only FC calibration in Section 2.4.

For sources at steps $t_a, t_b$, set $T = \max(t_a, t_b)$ and let $A(\theta)$ denote measured accuracy under a fixed protocol. The question is whether the constructed policy exceeds $\max_{\theta \in \mathcal{C}_T} A(\theta)$ without further RL or a larger inference budget. When the later source is the last evaluated checkpoint, this comparison covers the full measured history. Reasoning tokens count the full generated response, including

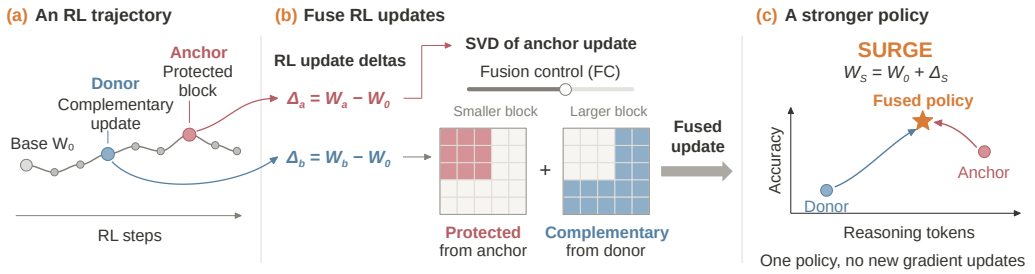



Figure 2: **SURGE overview.** Checkpoints from one RL trajectory are expressed as deltas before SVD. The anchor defines the protected update block; the donor supplies its complement. Recombination yields one policy without new gradient updates. The accuracy–token plot is schematic.

submitted code. We study improvements attainable from a fixed history, without claiming a fitted scaling law or improvement for every pair.

### 2.2 Delta-first singular decomposition

Figure 2 summarizes the construction. The anchor $\theta_a$ and donor $\theta_b$ share an architecture and RL initialization. The method is asymmetric: the anchor defines the protected block, so exchanging the inputs generally changes the output.

For each target matrix $W \in \mathbb{R}^{m\times n}$, we first form the updates relative to the start of RL:

$$\Delta_a = W_a - W_0, \qquad \Delta_b = W_b - W_0. \tag{2}$$

Only after this subtraction do we compute the economy singular value decomposition

$$\Delta_a = U\Sigma V^\top, \qquad r = \min(m,n), \qquad k = \lceil \mathrm{FC}\, r \rceil, \quad 0 < \mathrm{FC} \le 1. \tag{3}$$

The left and right singular vectors are eigenvectors of $\Delta_a\Delta_a^\top$ and $\Delta_a^\top\Delta_a$, respectively. Here *fusion control* (FC) sets the fraction of singular directions used to define a protected block. Applying SVD to $W_a$ instead would define a different geometry, dominated in part by the shared initialization. Delta-first construction makes the reference the change induced by RL itself. A component shared by the initialization and both checkpoints cancels before the spectral basis is computed. The resulting coordinates describe accumulated RL changes, rather than the absolute weight structure inherited before RL. Although $W_0$ cancels from the checkpoint difference $W_b - W_a$, it remains essential to this geometry.

### 2.3 Protected and complementary update components

Let $U_k, V_k$ contain the leading $k$ left and right singular vectors of $\Delta_a$. In matrix space with the Frobenius inner product, define

$$\mathcal{B}_{a,k} = \{U_k Z V_k^\top : Z \in \mathbb{R}^{k\times k}\},$$
$$\mathcal{K}_{a,k}(X) = (U_k U_k^\top) X (V_k V_k^\top), \qquad \mathcal{R}_{a,k}(X) = X - \mathcal{K}_{a,k}(X). \tag{4}$$

We call $\mathcal{B}_{a,k}$ the *protected update block*. The *protected-block projector* $\mathcal{K}_{a,k}$ extracts the retained component, while the *complement projector* $\mathcal{R}_{a,k}$ extracts the update outside that block. The index $a$ emphasizes that both operators are defined by the anchor's RL update. The protected object is a two-sided matrix block of dimension $k^2$, including interactions between the selected left and right directions, rather than only $k$ matched singular outer products.

SURGE preserves the anchor's protected component and admits the donor's complementary component:

$$\Delta_{\mathrm{S}} = \Delta_b + \mathcal{K}_{a,k}(\Delta_a - \Delta_b),$$
$$W_{\mathrm{S}} = W_0 + \Delta_{\mathrm{S}}. \tag{5}$$

The construction gives exact component identities, $\mathcal{K}_{a,k}(\Delta_{\mathrm{S}}) = \mathcal{K}_{a,k}(\Delta_a)$ and $\mathcal{R}_{a,k}(\Delta_{\mathrm{S}}) = \mathcal{R}_{a,k}(\Delta_b)$. These are statements about parameters. Because model computation is nonlinear, they do not imply independent preservation of the source policies' behaviors.

### 2.4 CALIBRATING FUSION CONTROL BY RETAINED UPDATE ENERGY

FC controls the protected block's size, not an interpolation amplitude. To relate this rank fraction to the anchor's update geometry, define the retained energy

$$\xi_a(\mathrm{FC}) = \frac{\sum_{\ell \in \mathcal{L}} \sum_{i=1}^{k_\ell} \sigma_{\ell,i}^2}{\sum_{\ell \in \mathcal{L}} \sum_{i=1}^{r_\ell} \sigma_{\ell,i}^2}, \qquad k_\ell = \lceil \mathrm{FC} r_\ell \rceil, \tag{6}$$

where $\mathcal{L}$ contains the target matrices. This is the fraction of anchor-update squared Frobenius norm preserved by the protected blocks. It depends on the anchor, not the donor. The entire curve follows from cumulative squared singular values of the same SVD, without additional decompositions or inference. FC is the rank coordinate used by the algorithm; $\xi_a$ calibrates what that coordinate preserves in the anchor's update. We use the discrete rule

$$\mathrm{FC}_{\mathrm{spec}} = \arg\min_{f \in \mathcal{F}} |\xi_a(f) - \tau|, \qquad \tau = 0.95, \quad \mathcal{F} = \{0.1, 0.2, \ldots, 0.9\}. \tag{7}$$

We identified $\tau = 0.95$ from the initial mathematical study and fixed it before applying the same procedure to OLMo. The common rule returns FC $= 0.8$, $0.8$, and $0.7$ for DeepSeek, Nemotron, and OLMo, respectively. Once the checkpoint pair and target are fixed, no accuracy sweep is required: cumulative singular values determine FC, and one fusion produces the policy. The target is empirical, not a guarantee of reward optimality; the rule selects the nearest candidate without requiring $\xi_a \geq \tau$. Section 3.4 analyzes the resulting policy family.

For a fixed projector, fusion is the closest donor update whose protected component matches the anchor. Appendix A gives the variational characterization, proof, and details of the native references.

### 2.5 WHY AN OPTIMIZATION PATH NEED NOT EXHAUST CAPABILITY

RL optimizes a coupled policy and realizes one sequence of joint parameter updates. It does not enumerate the recombinations of updates accumulated along that sequence. These are distinct spaces: the optimizer's visited checkpoints and the policies constructible from their components. Since policy behavior is nonlinear in the weights, scores of the latter need not lie between those of their sources. Section 3 tests this possibility against native trajectories and geometric controls.

A shorter-response donor may avoid some unproductive exploration or redundant continuation. If its useful updates complement the anchor, fusion can improve correctness as well as efficiency; the ablations test whether structured recombination matters beyond reducing token usage.

When both $\mathcal{K}_{a,k}(W_b - W_a)$ and $\mathcal{R}_{a,k}(W_b - W_a)$ are nonzero, the fusion in Equation 5 lies off the affine line through the two source matrices. It explores a recombination that no scalar interpolation coefficient can reproduce. We use "beyond the RL trajectory" in the performance sense: exceeding the measured native policies. Appendix A gives a conditional local reward interpretation.

**Implementation.** We fuse the query, key, value, and output attention projections and the gate, up, and down feed-forward projections, using one FC across layers. Embeddings, normalization parameters, language-model heads, and attention biases are copied from the anchor. Algorithm 1 states the tensor-level procedure. SVD and fusion arithmetic use FP32, and checkpoints are saved in BF16. The output runs as one policy. For the two math families, a full pass over 196 matrices, including spectral diagnostics and checkpoint writing, takes 37.2–40.1 seconds on one RTX A6000 with warm caches (Appendix G).

## 3 EXPERIMENTS

We evaluate whether a fixed RL history supports stronger policies than its measured checkpoints, whether this opportunity persists across scale and domain, and which update structure matters. Main tables evaluate one pair from each of three histories. Full-trajectory comparisons cover 19 native checkpoints each on DeepSeek AIME24 and OLMo HumanEval+, through their respective anchors.

Table 1: **Main results.** Accuracy (%) and mean reasoning tokens occupy separate labeled rows. Bold marks the best accuracy per family and benchmark. Avg. is the unweighted mean across benchmarks. Each Δ compares SURGE with its anchor: accuracy in percentage points and tokens as relative change; green indicates improvement and red indicates regression. Changes are computed before rounding.

| Policy | Metric | AIME24 | AIME25 | HMMT25 | BRUMO25 | CMIMC25 | AMC23 | MATH | Minerva | Olympiad | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **DeepSeek** | | | | | | | | | | | |
| Donor | Acc. | 47.50 | 35.83 | 19.17 | 46.67 | 21.88 | 88.13 | 90.75 | 50.64 | 64.80 | 51.71 |
| | Tokens | 7,651 | 6,912 | 7,574 | 7,194 | 8,525 | 4,753 | 3,115 | 4,582 | 5,570 | 6,208 |
| Anchor | Acc. | 50.83 | 36.67 | 20.00 | 49.17 | 22.50 | 90.00 | 91.60 | 52.02 | 66.58 | 53.26 |
| | Tokens | 8,065 | 7,927 | 8,109 | 7,892 | 9,147 | 5,472 | 3,656 | 4,796 | 6,246 | 6,812 |
| SURGE | Acc. | **54.17** | **38.33** | **23.33** | **51.67** | **25.63** | **92.50** | **92.40** | **52.76** | **67.32** | **55.35** |
| | Tokens | 7,798 | 7,383 | 7,693 | 7,285 | 8,700 | 5,225 | 3,415 | 4,742 | 5,832 | 6,453 |
| Δ | Acc. | +3.34 | +1.66 | +3.33 | +2.50 | +3.13 | +2.50 | +0.80 | +0.74 | +0.74 | +2.08 |
| | Tokens | -3.3% | -6.9% | -5.1% | -7.7% | -4.9% | -4.5% | -6.6% | -1.1% | -6.6% | -5.3% |
| **Nemotron** | | | | | | | | | | | |
| Donor | Acc. | 60.00 | 50.83 | 35.00 | 61.67 | 29.38 | 91.88 | 92.45 | 33.18 | 73.81 | 58.69 |
| | Tokens | 9,212 | 10,387 | 10,652 | 8,844 | 10,847 | 5,687 | 3,754 | 6,142 | 6,536 | 8,007 |
| Anchor | Acc. | 68.33 | 60.83 | 35.83 | 65.83 | 38.75 | 95.00 | 94.25 | 33.55 | **77.26** | 63.29 |
| | Tokens | 13,030 | 14,395 | 14,841 | 12,548 | 15,603 | 8,393 | 5,452 | 8,627 | 9,396 | 11,365 |
| SURGE | Acc. | **70.00** | **63.33** | **39.17** | **68.33** | **41.88** | **97.50** | **94.35** | **34.28** | 76.85 | **65.08** |
| | Tokens | 11,395 | 12,294 | 13,071 | 11,061 | 13,223 | 7,375 | 4,911 | 7,513 | 8,348 | 9,910 |
| Δ | Acc. | +1.67 | +2.50 | +3.34 | +2.50 | +3.13 | +2.50 | +0.10 | +0.73 | -0.41 | +1.78 |
| | Tokens | -12.5% | -14.6% | -11.9% | -11.9% | -15.3% | -12.1% | -9.9% | -12.9% | -11.2% | -12.8% |

### 3.1 EXPERIMENTAL SETUP

**RL training follows JustRL.** The two math trajectories use DeepSeek-R1-Distill-Qwen-1.5B and OpenMath-Nemotron-1.5B, following JustRL's single-stage GRPO recipe (He et al., 2025; DeepSeek-AI, 2025; Moshkov et al., 2025). Training uses veRL, DAPO-Math-17k, and binary correctness rewards from the DAPO rule-based verifier (Shao et al., 2024; Sheng et al., 2025; Yu et al., 2025). Each prompt produces eight rollouts at temperature 1.0; group-relative rewards drive updates with learning rate $10^{-6}$ and batch size 256. The recipe uses no explicit length penalty, KL loss, entropy regularization, or stage changes. Appendix A provides the full settings.

**Benchmarks and evaluation.** We use JustRL's nine mathematical benchmarks: AIME24/25, HMMT25, BRUMO25, CMIMC25, AMC23, MATH-500 (MATH), Minerva, and OlympiadBench (Olympiad) (LI et al., 2024; Balunovic et al., 2025; Hendrycks et al., 2021; Lewkowycz et al., 2022; He et al., 2024). Accuracy averages correctness over four responses per problem; reasoning tokens are averaged over generated responses. Decoding follows JustRL with temperature 0.7, top-$p$ 0.9, and a 32K response limit. Scoring combines rule-based evaluation and CompassVerifier (He et al., 2025). These are sampled single-response accuracies, not best-of-four success rates. Avg. in Table 1 weights benchmarks equally for both accuracy and tokens. Appendix C reports DeepSeek accuracy variation across the four evaluation rounds.

**Evaluated configurations.** The math experiments use DeepSeek steps 3600/2600 and Nemotron steps 3440/1200 as anchor/donor, respectively. The anchors have strong aggregate accuracy; the donors offer shorter responses while retaining competitive accuracy in the available source-policy evaluations. Equation 7 returns FC $= 0.8$ for both. Each comparison holds the RL history, architecture, and decoding protocol fixed.

### 3.2 SCALING A FIXED MATHEMATICAL-REASONING HISTORY

SURGE raises DeepSeek's measured AIME24 accuracy ceiling from 50.83% to 54.17% at the same RL horizon (Figure 1). It exceeds all 19 native measurements while using 7,798 tokens versus the

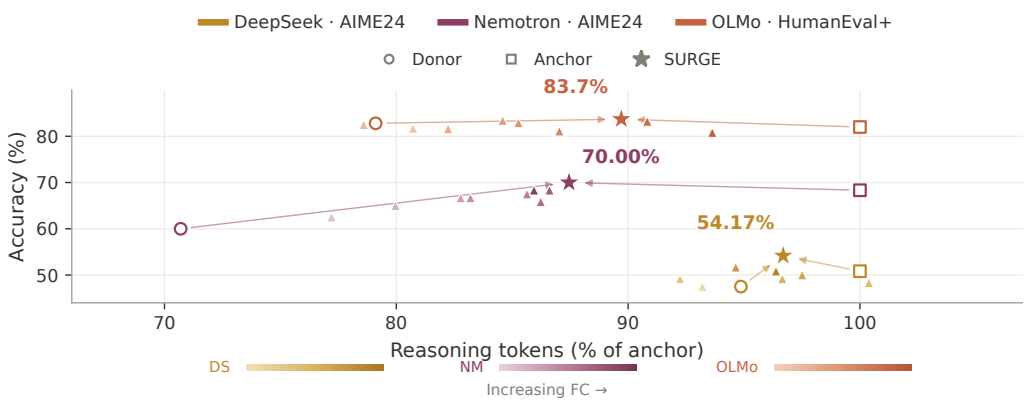



Figure 3: **New policies from fixed RL histories.** Tokens are normalized to each anchor. Triangles are measured FC candidates; stars mark the spectrally calibrated policies. Arrows connect source checkpoints to the resulting policy. All candidates use four rounds; comparisons are within each family and benchmark.

anchor's 8,065. The benefit extends across the evaluation suite: accuracy improves and reasoning tokens fall on all nine benchmarks. Avg. accuracy rises from 53.26% to 55.35% (+2.08 points), with 5.3% fewer tokens.

Nemotron's Avg. rises from 63.29% to 65.08% (+1.78 points), with 12.8% fewer tokens. Accuracy improves on eight of nine benchmarks; only Olympiad decreases, by 0.41 points. Reasoning tokens fall on all nine. AIME24 reaches 70.00%, above the anchor's 68.33% and donor's 60.00%, with 12.5% fewer tokens than the anchor.

Across all three histories in Figure 3, the spectrally calibrated policies exceed both sources on the displayed benchmark, with fewer tokens than the anchor under the same response budget.

### 3.3 Cross-scale and cross-domain validation: 7B code

Olmo-3.1-7B-RL-Zero-Code changes the scale, architecture, and reasoning domain. Training starts from Olmo-3-1025-7B and follows OlmoRL on Dolci-RL-Zero-Code-7B: GRPO-based execution rewards, eight responses per prompt, learning rate $10^{-6}$, and a 16K response limit (Ettinger et al., 2025). HumanEval+ and MBPP+ use EvalPlus (Liu et al., 2023), with four rounds of one response per problem at temperature 1.0, top-$p$ 1.0, and a 32K limit. Appendix H gives the full protocol and 19 native checkpoint measurements.

Step 1950 has the highest two-benchmark mean among six jointly evaluated checkpoints and serves as the anchor. The shorter-response donor is step 1500. Equation 7 selects FC $= 0.7$ from the anchor spectrum with the fixed target.

Table 2 reports 83.7% HumanEval+ accuracy: 1.7 points above the anchor and 0.9 above the best native checkpoint, the donor. MBPP+ matches the anchor at reported precision. Tokens fall by 10.3% and 10.5% relative to the anchor. These point estimates extend the observed gain beyond native checkpoint selection to 7B code.

**A concrete policy difference.** Table 3 shows a case where SURGE more consistently follows the program specification. Direct set equality appears in three of its four responses, versus one for the donor and none for the anchor. This selected example illustrates a behavioral difference, rather than estimating its prevalence; Appendix E provides additional cases.

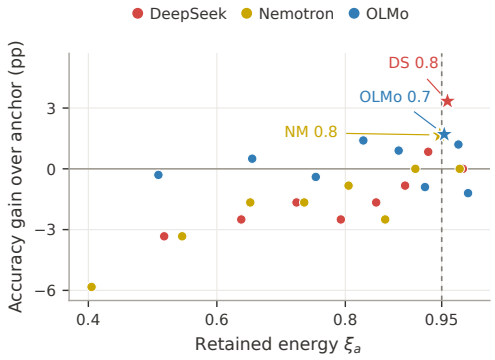



Figure 4: **Shared spectral calibration across trajectories.** Measured candidates (dots) and selected FC settings (stars); dashed line: $\xi_a = 0.95$. AIME24 for math, HumanEval+ for code.

Table 2: **7B coding results.** Accuracy (%) and reasoning tokens (k). Δ gives changes from the anchor in points and percent.

| Policy | Metric | HumanEval+ | MBPP+ |
|---|---|---|---|
| Donor | Acc. (%) | 82.8 | 68.4 |
| | Tokens (k) | 4.17 | 3.02 |
| Anchor | Acc. (%) | 82.0 | **71.2** |
| | Tokens (k) | 5.28 | 3.86 |
| **SURGE** | Acc. (%) | **83.7** | **71.2** |
| | Tokens (k) | 4.73 | 3.45 |
| Δ | Accuracy | +1.7 | 0.0 |
| | Tokens | −10.3% | −10.5% |

Table 3: **Specification fidelity on HumanEval+ #54.** A selected OLMo coding example. Correctness counts are over four responses per policy, each checked against the original and additional EvalPlus tests.

**Task.** Implement `same_chars(s0, s1)`: compare the two strings' character sets. Order and repetition do not matter; case matters.

| **Donor** | **Anchor** | **SURGE** |
|---|---|---|
| One response compares sets correctly. The others compare sorted characters, lowercase the strings, or compare only set sizes. | Two responses count character multiplicities; two compare sets after lowercasing. These reject valid repetitions or erase case distinctions. | Three responses compare sets directly, ignoring repetition while preserving case. One response still sorts lowercased strings. |
| Observed correct: **1/4** | Observed correct: **0/4** | Observed correct: **3/4** |

**Passing implementation:** `return set(s0) == set(s1)`

### 3.4 A SHARED SPECTRAL COORDINATE ACROSS TRAJECTORIES

Figure 4 compares accuracy gains on the common retained-energy axis. At the fixed target, the weight-only rule selects strong policies across all three trajectories despite their different rank fractions. Matrix shapes and energy distributions explain why matching preserved update energy is more consistent than imposing the same FC (Appendix F). The sweeps characterize sensitivity; a new pair does not require an accuracy sweep once the target is fixed.

The calibrated settings coincide with the highest measured accuracies in all three sweeps. On HumanEval+, the neighboring FC $= 0.8$ reaches 83.2%, versus 83.7% for the rule's FC $= 0.7$; these small differences do not establish a statistically resolved optimum. Other useful policies exist in the same history: FC $= 0.4$ reaches 83.4% with 4,462 tokens.

Accuracy is non-monotonic although retained energy increases with FC: parameter-space motion does not linearly interpolate policy quality. DeepSeek at FC $= 0.7$ reaches 51.67% with 7,633 tokens, while 0.8 adds 2.50 points for 165 tokens. The broader family therefore supports different deployment choices; spectral calibration supplies one reproducible policy without evaluating them all. Full raw-FC measurements are in Appendix D.

### 3.5 ABLATIONS: WHICH UPDATE GEOMETRY MATTERS?

Table 4 compares reuse of the same DeepSeek history. Scalar and spectral controls act on 196 body linear matrices; checkpoint averages include all parameters.

Table 4: **DeepSeek checkpoint-reuse controls on AIME24.** Four evaluation rounds per policy. Scalar and spectral controls use the same source pair; late averaging uses its RL history. The random row averages four constructions. Full scalar and averaging results are in Appendix B.

| Method | Accuracy (%) | Reasoning tokens |
|---|---|---|
| Anchor | 50.83 | 8,065 |
| Donor | 47.50 | 7,651 |
| Best tested non-endpoint scalar ($\alpha = -0.5$) | 49.17 | 9,141 |
| Linear interpolation ($\alpha = 0.71$) | 47.50 | 7,659 |
| Late checkpoint average (3 checkpoints) | 50.83 | 8,078 |
| Anchor-only spectral truncation | 44.17 | 8,320 |
| Absolute-weight SVD | 41.67 | 8,173 |
| Random subspace (mean of 4) | 47.29 | 7,574 |
| Trailing singular subspace | 44.17 | 7,743 |
| SURGE (Delta-SVD) | **54.17** | 7,798 |

Replacing delta-first SVD with absolute-weight SVD reduces AIME24 accuracy from 54.17% to 41.67%, with lower accuracy on all nine benchmarks (Appendix B). At the same FC and block size, the absolute-weight block captures 28.5% of RL-update energy, versus 95.9% for the delta-defined block (Appendix F).

**The donor contributes beyond anchor truncation.** Retaining only the anchor's protected update, $W_{\mathrm{trunc}} = W_0 + \mathcal{K}_{a,k}(\Delta_a)$, uses the same projector and FC $= 0.8$. It reaches 44.17% accuracy with 8,320 tokens, versus SURGE's 54.17% and 7,798. Removing the anchor tail alone does not recover the gain: adding the donor's complementary update raises measured accuracy by 10.00 points in this pair.

**Scalar combinations and checkpoint averaging.** For $W_\alpha = (1-\alpha)W_a + \alpha W_b$, six tested interpolation and extrapolation coefficients reach at most 49.17% ($\alpha = -0.5$), below the anchor and SURGE. To control for displacement magnitude, define

$$\eta = \frac{\left(\sum_{\ell \in \mathcal{L}} \|W_{\mathrm{S}}^{\ell} - W_a^{\ell}\|_F^2\right)^{1/2}}{\left(\sum_{\ell \in \mathcal{L}} \|W_b^{\ell} - W_a^{\ell}\|_F^2\right)^{1/2}} = 0.713, \tag{8}$$

Scalar combinations have normalized distance $|\alpha|$. Thus $\alpha = 0.71$ approximately matches SURGE's aggregate distance, yet reaches only 47.50%; direction matters beyond magnitude. Three- and six-checkpoint late averages reach 50.83% and 49.17%, respectively. Neither tested average exceeds the anchor. Appendix B gives the full grid and averaging windows.

**Protected geometry beyond token reduction.** At fixed block size, random left/right orthonormal bases yield 47.29% mean accuracy across four independent constructions. Retaining the last $k$ singular directions instead yields 44.17%. Both use fewer tokens than SURGE (7,574 and 7,743 versus 7,798), while losing 6.88 and 10.00 accuracy points. These controls favor the leading RL-update block over the tested alternatives; neither block size nor shorter reasoning alone recovers the gain.

Across the repeated math evaluations, within-cell accuracy and reasoning tokens have a weak negative association ($r = -0.181$, permutation $p = 0.021$), consistent with but not establishing a causal efficiency benefit (Appendix C.1).

**Implications for policy-space scaling.** The controls distinguish expanding the deployable policy set from merely perturbing weights or shortening responses. This expansion spends offline construction compute while reusing the completed RL history. Once a checkpoint pair and retention target are fixed, the anchor spectrum determines one policy; evaluating the broader family is optional and has a separate cost. The reported sweeps diagnose that family, whereas deployment uses a single checkpoint under the original decoding protocol. Thus, the additional scaling opportunity lies in how already-computed RL updates are used.

## 4 Related Work

**Scaling reasoning with verifiable rewards.** DeepSeekMath introduced GRPO for mathematical reasoning, and DeepSeek-R1 demonstrated that large-scale RL can elicit effective reasoning behavior from correctness rewards (Shao et al., 2024; DeepSeek-AI, 2025). DAPO develops an open RL system with dynamic sampling and changes to clipping and token-level optimization (Yu et al., 2025). ProRL studies prolonged training, using KL control and reference-policy resets to sustain exploration and expand reasoning performance (Liu et al., 2025a). JustRL emphasizes a fixed, single-stage recipe for small reasoning models (He et al., 2025); we follow its setting and study how to reuse its accumulated updates. Efficiency is also a property of the training objective: Dr. GRPO identifies length-related bias in GRPO and shows that correcting it can improve both performance and token efficiency (Liu et al., 2025b). At inference time, s1 controls reasoning compute through budget forcing (Muennighoff et al., 2025). SURGE complements these training- and inference-time approaches by constructing one deployable policy from stored checkpoints, without further policy-gradient updates or a larger decoding budget.

**Model merging and update composition.** Model soups average fine-tuned weights without inference-time ensembling, while Fisher-weighted averaging accounts for parameter importance (Wortsman et al., 2022; Matena & Raffel, 2022). Task arithmetic composes updates relative to a common initialization; TIES-Merging reduces sign interference, and DARE sparsifies and rescales updates (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024). Spectral methods address their matrix structure: Task Singular Vectors reduces interference among task-update directions, and isotropic merging rebalances spectra using common and task-specific subspaces (Gargiulo et al., 2025; Marczak et al., 2025). More directly, ResMerge finds useful information in both the leading heads and residuals of RL task vectors, combining multiple RL experts through residual consensus and agreement-gated head correction (Sun et al., 2026). These studies establish tools for update composition. Our central question concerns the capability accessible from one RL training history. SURGE uses spectral composition to expand that history's candidate policy space, and evaluates whether it exceeds the measured native trajectory. These experiments examine policy-space scaling from already-computed RL updates across domains and through geometric controls.

**Reusing RL training trajectories.** AlphaRL predicts dominant update components from early checkpoints through accuracy-conditioned regression (Cai et al., 2025). RELEX decomposes stacked checkpoint deltas and linearly extrapolates their rank-1 trajectory coefficients (Wei et al., 2026). NExt trains a nonlinear predictor of rank-1 update representations from LoRA-based RLVR trajectories (Chen et al., 2026). These methods forecast additional policies from observed training history. Extrapolative weight averaging also extends correctness–efficiency frontiers in code RL, combining policies trained under different unit-test coverage (Zheng et al., 2026). There, efficiency concerns generated programs under execution resource limits; here, we measure reasoning tokens. SURGE expands the deployable policy set by recombining components of observed RL updates, with a weight-only calibration rule after source selection. It uses the history's update geometry without fitting a trajectory predictor or specifying a future training step. This offers a complementary construction for policy-space scaling from an existing RL budget.

## 5 Conclusion

An RL run can deliver more capability than its best measured checkpoint. Across 1.5B math and 7B code, one weight-only calibration procedure constructs policies that improve on their sources; on the evaluated DeepSeek and OLMo trajectories, they exceed every measured native checkpoint. The optimizer's path therefore need not define the capability ceiling of its training history. Policy-space scaling turns that history into a reusable policy space, complementing new optimization and additional inference with offline reuse of already-computed updates. Its construction and evaluation costs are explicit, while deployment remains a single policy. A broader RL scaling strategy should account both for how updates are produced and for how their accumulated history is used: additional capability can reside in policies the optimizer never visited.

### AI USE STATEMENT

Generative AI assistants (ChatGPT and Codex) were used to help refine the conceptual framing and hypotheses, discuss methodology and experimental controls, formulate and check mathematical arguments, interpret quantitative results, and summarize qualitative reasoning cases. They also assisted with literature search and comparison, manuscript drafting and editing, analysis and visualization code, and figure preparation. During revision, numerical summaries were cross-checked against the experiment records, references were consulted at their original sources, and the compiled figures and manuscript were inspected. The authors take responsibility for the final content, including all AI-assisted text, claims, code, and figures.

### ETHICS STATEMENT

This study concerns existing language models and public mathematical reasoning and coding benchmarks. Policies constructed from an RL history may inherit biases and unsafe behaviors from their source models and training data. The reported gains concern task accuracy and reasoning-token cost; they do not establish deployment safety or mitigation of these inherited risks.

### REPRODUCIBILITY STATEMENT

Section 2 specifies the policy construction and weight-only calibration rule, and Algorithm 1 gives the complete procedure. Section 3 and Appendices A and H document the source checkpoints, training recipes, benchmarks, decoding settings, scoring procedures, and numerical precision. Appendices B–D provide the additional controls, evaluation variability, and controller measurements; Appendix F details the retained-energy calibration, and Appendix G documents the compute-cost measurements and their scope. Appendix A also states the assumptions and proof for the projection interpretation.

## REFERENCES

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanovic, and Martin T. Vechev. Matharena: Evaluating llms on uncontaminated math competitions. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/1d27c01ebd3e3aebe226b44fc970d803-Abstract-Datasets_and_Benchmarks_Track.html.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guiquan Liu, and Junfeng Fang. On predictability of reinforcement learning dynamics for large language models. *CoRR*, abs/2510.00553, 2025. doi: 10.48550/ARXIV.2510.00553. URL https://doi.org/10.48550/arXiv.2510.00553.

Zhipeng Chen, Tao Qian, Wayne Xin Zhao, and Ji-Rong Wen. Low-rank optimization trajectories modeling for LLM RLVR acceleration. *CoRR*, abs/2604.11446, 2026. doi: 10.48550/ARXIV.2604.11446. URL https://doi.org/10.48550/arXiv.2604.11446.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee F. Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester

James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, William Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3. *CoRR*, abs/2512.13961, 2025. doi: 10.48550/ARXIV.2512.13961. URL https://doi.org/10.48550/arXiv.2512.13961.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18695–18705. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01742. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gargiulo_Task_Singular_Vectors_Reducing_Task_Interference_in_Model_Merging_CVPR_2025_paper.html.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. Justrl: Scaling a 1.5b LLM with a simple RL recipe. *CoRR*, abs/2512.16649, 2025. doi: 10.48550/ARXIV.2512.16649. URL https://doi.org/10.48550/arXiv.2512.16649.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.211. URL https://doi.org/10.18653/v1/2024.acl-long.211.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural*

*Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/1a22b912945fb7c0bdd079e792b31b6f-Abstract-Conference.html.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025b. doi: 10.48550/ARXIV.2503.20783. URL https://doi.org/10.48550/arXiv.2503.20783.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartlomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/marczak25a.html.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. AIMO-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *CoRR*, abs/2504.16891, 2025. doi: 10.48550/ARXIV.2504.16891. URL https://doi.org/10.48550/arXiv.2504.16891.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1025. URL https://doi.org/10.18653/v1/2025.emnlp-main.1025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Yandu Sun, ZhiYan Hou, Haokai Ma, Yuheng Jia, Junfeng Fang, Haiyun Guo, Hongyan An, Weizhen Wang, and Jinqiao Wang. Resmerge: Residual-based spectral merging of large language models. *CoRR*, abs/2606.02252, 2026. doi: 10.48550/ARXIV.2606.02252. URL https://doi.org/10.48550/arXiv.2606.02252.

Zhepei Wei, Xinyu Zhu, Wei-Lin Chen, Chengsong Huang, Jiaxin Huang, and Yu Meng. You only need minimal RLVR training: Extrapolating llms via rank-1 trajectories. *CoRR*, abs/2605.21468, 2026. doi: 10.48550/ARXIV.2605.21468. URL https://doi.org/10.48550/arXiv.2605.21468.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Kunhao Zheng, Pierre Chambon, Juliette Decugis, Jonas Gehring, Taco Cohen, Benjamin Négrevergne, and Gabriel Synnaeve. Extrapolative weight averaging reveals correctness-efficiency frontiers in code RL. *CoRR*, abs/2605.28751, 2026. doi: 10.48550/ARXIV.2605.28751. URL https://doi.org/10.48550/arXiv.2605.28751.

# A TRAINING, EVALUATION, AND FUSION DETAILS

Table 5 records the math training and evaluation settings; Appendix H gives the code settings. All main-table results use four responses per question, unlike JustRL's higher sample count on several tasks. The qualitative examples in Appendix E are per-problem diagnostics.

Table 5: **Math configuration.** Training settings follow JustRL; evaluation settings and four-response main-table estimates are shared by each fusion and its source checkpoints.

| Stage | Setting | Value |
|---|---|---|
| RL | Algorithm / framework | GRPO / veRL |
| | Training data | DAPO-Math-17k |
| | Reward | Binary correctness, DAPO rule-based verifier |
| | Learning rate | $10^{-6}$, constant |
| | Train batch / PPO mini-batch | 256 / 64 |
| | Micro-batch per GPU | 1 |
| | Responses per prompt / temperature | 8 / 1.0 |
| | Clip-ratio interval | $[0.8, 1.28]$ |
| | Maximum prompt / response | 1K / 15K tokens |
| | KL loss / entropy regularization | None / none |
| | Length penalty / stage schedule | None / single stage |
| Evaluation | Temperature / top-$p$ | 0.7 / 0.9 |
| | Maximum response | 32K tokens |
| | Main-table responses per problem | 4 for every benchmark |
| | Verification | Rule-based scoring plus CompassVerifier |
| Fusion | Primary FC | 0.8 for both math families ($\xi_a$ nearest 0.95) |
| | Reference for SVD | Anchor update $W_a - W_0$ |
| | Computation / saved precision | FP32 / BF16 |
| | Non-target parameters | Copied from anchor |

**Algorithm 1** SURGE: weight-only calibration and policy construction

**Require:** Initialization $\theta_0$, anchor $\theta_a$, donor $\theta_b$, target matrices $\mathcal{L}$, grid $\mathcal{F}$ and target $\tau$ from Equation 7

**for** each matrix $\ell \in \mathcal{L}$ **do**
  $\Delta_a^\ell \leftarrow W_a^\ell - W_0^\ell$, $\Delta_b^\ell \leftarrow W_b^\ell - W_0^\ell$
  Cache $U^\ell, \Sigma^\ell, (V^\ell)^\top \leftarrow \mathrm{SVD}(\Delta_a^\ell)$
**end for**
Compute $\xi_a(f)$ for $f \in \mathcal{F}$ from cached singular values (Equation 6)
$\mathrm{FC}_{\mathrm{spec}} \leftarrow \arg\min_{f \in \mathcal{F}} |\xi_a(f) - \tau|$ ▷ No policy evaluation
$\theta_{\mathrm{S}} \leftarrow \theta_a$ ▷ Copy non-target parameters from the anchor
**for** each matrix $\ell \in \mathcal{L}$ of shape $m \times n$ **do**
  $k \leftarrow \max(1, \lceil \mathrm{FC}_{\mathrm{spec}} \min(m, n) \rceil)$; $U_k \leftarrow U^\ell_{:,1:k}$, $V_k \leftarrow V^\ell_{:,1:k}$
  $\Delta_{\mathrm{prot}} \leftarrow U_k(U_k^\top \Delta_a^\ell V_k)V_k^\top$ ▷ Protected anchor component
  $\Delta_{\mathrm{comp}} \leftarrow \Delta_b^\ell - U_k(U_k^\top \Delta_b^\ell V_k)V_k^\top$ ▷ Complementary donor component
  $W_{\mathrm{S}}^\ell \leftarrow W_0^\ell + \Delta_{\mathrm{prot}} + \Delta_{\mathrm{comp}}$
**end for**
**return** $\theta_{\mathrm{S}}$

**Computational cost.** For $W \in \mathbb{R}^{m \times n}$ and $r = \min(m, n)$, a dense economy SVD costs on the order of $mnr$ operations. Evaluating $U_k(U_k^\top X V_k)V_k^\top$ avoids materializing an $(mn) \times (mn)$ projector. The resulting checkpoint has the same tensor shapes as its sources. The SVD can be reused for different FC values and for the retained-energy curve in Equation 6. Appendix G reports measured fusion timings and a separate reconstruction of training-step cost.

**Boundary references.** The entries at 0 and 1 are native checkpoint references, not an endpoint-equivalence guarantee. At zero, the code retains at least one singular direction and copies anchor non-target parameters. At one, a rectangular matrix's full economy-SVD block need not be the

identity on the entire matrix space. The interior settings and delta-first projector define the analyzed method.

**A conditional local reward criterion.** Let $J(\theta)$ be a smooth expected-reward objective, with an $L$-Lipschitz gradient on the segment from the anchor to the fused parameters, and let $h = \theta_{\mathrm{S}} - \theta_a$. The standard smoothness bound gives

$$J(\theta_a + h) - J(\theta_a) \geq \nabla J(\theta_a)^\top h - \tfrac{L}{2}\|h\|_2^2. \tag{9}$$

If the directional benefit exceeds the curvature term, fusion improves the anchor; if the anchor is at least as strong as the donor under $J$, it exceeds both. This is a sufficient condition for a smooth objective, not a guarantee for sampled benchmark accuracy or discontinuous decoding rules. SURGE does not measure the gradient or assume that its projector enforces this condition.

**Projection interpretation.** The fused update is the unique closest donor update whose protected component matches the anchor:

$$\Delta_{\mathrm{S}} = \arg\min_X \|X - \Delta_b\|_F^2 \quad \text{subject to} \quad \mathcal{K}_{a,k}(X) = \mathcal{K}_{a,k}(\Delta_a). \tag{10}$$

*Proof.* The projectors satisfy $\mathcal{K}_{a,k}^2 = \mathcal{K}_{a,k}$, $\mathcal{R}_{a,k}^2 = \mathcal{R}_{a,k}$, and $\mathcal{K}_{a,k}\mathcal{R}_{a,k} = 0$. Their ranges are perpendicular under the Frobenius inner product, so the objective decomposes into $\|\mathcal{K}_{a,k}(X - \Delta_b)\|_F^2 + \|\mathcal{R}_{a,k}(X - \Delta_b)\|_F^2$. The constraint fixes the first term. The second has minimum zero, attained by setting $\mathcal{R}_{a,k}(X) = \mathcal{R}_{a,k}(\Delta_b)$. Together these conditions uniquely give Equation 5. □

## B ADDITIONAL ABLATION RESULTS

The ablations examine the decomposition reference, donor complement, displacement direction, protected block, and checkpoint averaging. Scalar and spectral variants use the DeepSeek sources in Section 3, target the same 196 body linear matrices, and keep non-target tensors from the anchor. Projector-based variants use FC $= 0.8$ and the same matrix-specific $k$. Checkpoint averages instead combine all parameters over the specified windows.

Table 6: **Expanded DeepSeek reuse controls on AIME24.** $\eta$ is normalized Frobenius distance over body linear matrices; Cap is the response-limit truncation rate. Distances are nonnegative, including for negative $\alpha$. A dash denotes an unreported distance.

| Variant | $\eta$ | Accuracy (%) | Reasoning tokens | Cap (%) |
|---|---|---|---|---|
| Anchor-only truncation | 0.329 | 44.17 | 8,320 | 2.5 |
| Scalar $\alpha = -0.50$ | 0.500 | 49.17 | 9,141 | 3.3 |
| Scalar $\alpha = -0.25$ | 0.250 | 47.50 | 7,723 | 0.0 |
| Scalar $\alpha = 0.25$ | 0.250 | 45.00 | 8,521 | 1.7 |
| Scalar $\alpha = 0.50$ | 0.500 | 44.17 | 8,028 | 0.8 |
| Scalar $\alpha = 0.71$ | 0.710 | 47.50 | 7,659 | 1.7 |
| Scalar $\alpha = 0.75$ | 0.750 | 46.67 | 7,275 | 0.8 |
| Average: 3 checkpoints | — | 50.83 | 8,078 | 0.8 |
| Average: 6 checkpoints | — | 49.17 | 7,977 | 2.5 |
| Trailing singular block | 0.834 | 44.17 | 7,743 | 0.8 |
| Absolute-weight SVD | 0.881 | 41.67 | 8,173 | 2.5 |

**Truncation, scalar combinations, and averaging.** Anchor-only truncation changes the anchor by $-\mathcal{R}_{a,k}(\Delta_a)$, toward initialization in the complementary block, with normalized distance 0.329. For scalar combinations, distance is $|\alpha|$; negative coefficients extrapolate away from the donor. The best measured grid value is descriptive, not an optimum over the continuous line. The three-checkpoint average uses steps 3200/3400/3600; the six-checkpoint average uses 2600/2800/3000/3200/3400/3600. Both uniformly average all parameters. Each policy in Table 6 uses four rounds under the main AIME24 protocol.

Table 7: **Absolute-weight versus delta-first SVD on DeepSeek.** Only the matrix used to define the singular basis changes; FC and the fusion rule are held fixed. Cells report accuracy (%) / mean reasoning tokens.

| Benchmark | Delta-SVD | Absolute-weight SVD |
|---|---|---|
| AIME24 | **54.17** / 7,798 | 41.67 / 8,173 |
| AIME25 | **38.33** / 7,383 | 35.00 / 7,368 |
| HMMT25 | **23.33** / 7,693 | 19.17 / 7,331 |
| BRUMO25 | **51.67** / 7,285 | 50.83 / 6,941 |
| CMIMC25 | **25.63** / 8,700 | 23.75 / 8,415 |
| AMC23 | **92.50** / 5,225 | 88.75 / 4,947 |
| MATH | **92.40** / 3,415 | 91.00 / 3,199 |
| Minerva | **52.76** / 4,742 | 52.11 / 4,493 |
| Olympiad | **67.32** / 5,832 | 66.58 / 5,639 |

The absolute-weight variant changes only the SVD input from $W_a - W_0$ to $W_a$. The remainder of the fusion rule is identical. Delta-first SVD gives higher accuracy on all nine benchmarks, although it does not uniformly use fewer reasoning tokens than the absolute-weight variant. This comparison concerns the geometry that preserves useful RL changes; it does not assume that lower token usage by itself implies a stronger reasoning policy.

Table 8: **Random-subspace constructions on DeepSeek AIME24.** Each repetition independently samples random left and right orthonormal bases for every target matrix at the same block size as SURGE.

| Construction | Accuracy (%) | Reasoning tokens |
|---|---|---|
| 1 | 50.83 | 7,239 |
| 2 | 45.83 | 8,182 |
| 3 | 47.50 | 7,205 |
| 4 | 45.00 | 7,668 |
| Mean | 47.29 | 7,574 |

For each repetition and target matrix of shape $m \times n$, sample Gaussian matrices of shapes $m \times k$ and $n \times k$, and take their QR factors as the left and right orthonormal bases. Each target matrix receives independently sampled bases. Four independent constructions produce the results above; they are distinct from the four sampled responses per evaluation problem. The trailing-subspace control instead uses the last $k$ singular vectors from the same delta SVD. The displacement ratio in Equation 8 concatenates all target-matrix changes; $\alpha = 0.71$ is an approximate, rounded match to the measured ratio 0.713.

# C EVALUATION VARIABILITY

Table 9: **DeepSeek variability across four evaluation rounds.** Entries give accuracy mean ± standard deviation / variance, computed across four round-level accuracy estimates. These are evaluation-sampling statistics, not variability across independent RL training runs.

| Benchmark | SURGE | Anchor | Donor |
|---|---|---|---|
| AIME24 | $\mathbf{54.17} \pm 6.40$ / 40.97 | $50.83 \pm 7.95$ / 63.19 | $47.50 \pm 5.95$ / 35.42 |
| AIME25 | $\mathbf{38.33} \pm 2.89$ / 8.33 | $36.67 \pm 4.71$ / 22.22 | $35.83 \pm 2.76$ / 7.64 |
| HMMT25 | $\mathbf{23.33} \pm 5.77$ / 33.33 | $20.00 \pm 3.33$ / 11.11 | $19.17 \pm 1.44$ / 2.08 |
| BRUMO25 | $\mathbf{51.67} \pm 1.67$ / 2.78 | $49.17 \pm 1.44$ / 2.08 | $46.67 \pm 3.33$ / 11.11 |
| CMIMC25 | $\mathbf{25.63} \pm 2.07$ / 4.30 | $22.50 \pm 4.33$ / 18.75 | $21.88 \pm 2.72$ / 7.42 |
| AMC23 | $\mathbf{92.50} \pm 3.06$ / 9.38 | $90.00 \pm 2.50$ / 6.25 | $88.13 \pm 3.70$ / 13.67 |
| MATH | $\mathbf{92.40} \pm 0.58$ / 0.34 | $91.60 \pm 0.14$ / 0.02 | $90.75 \pm 0.48$ / 0.23 |
| Minerva | $\mathbf{52.76} \pm 1.73$ / 3.01 | $52.02 \pm 0.41$ / 0.17 | $50.64 \pm 1.65$ / 2.73 |
| Olympiad | $\mathbf{67.32} \pm 0.46$ / 0.21 | $66.58 \pm 0.16$ / 0.03 | $64.80 \pm 0.74$ / 0.54 |

The four round-level scores correspond to sampling one response per question in each round. Their mean is the reported four-response accuracy. Standard deviations and variances quantify variability in these round-level estimates, not variation across independent RL training runs. We report descriptive accuracy differences and do not claim per-benchmark statistical significance. The table retains rounded statistics, so squaring a displayed standard deviation need not reproduce the displayed variance exactly. These statistics describe individual cells; the pooled association across all 54 cells is analyzed below.

## C.1 WITHIN-CELL ACCURACY–TOKEN CORRELATION

The corrected run-level report contains four paired accuracy/token observations for each of 54 family–policy–benchmark cells. Let $a_{ci}$ and $t_{ci}$ be the reported accuracy and mean reasoning tokens in round $i \in \{1, \ldots, 4\}$ of cell $c$. We standardize both quantities within the cell:

$$z^a_{ci} = \frac{a_{ci} - \bar{a}_c}{s^a_c}, \qquad z^t_{ci} = \frac{t_{ci} - \bar{t}_c}{s^t_c}, \qquad r = \frac{\sum_{c=1}^{54} \sum_{i=1}^{4} z^a_{ci} z^t_{ci}}{3 \cdot 54}, \tag{11}$$

where each $s_c$ is the sample standard deviation across the four rounds. All cells have nonzero variation in both metrics. Because each standardized vector has squared norm 3, Equation 11 is both the pooled Pearson correlation and the equal-weight mean of the 54 cell-level correlations. Scaling, in addition to centering, prevents high-variance benchmarks from dominating the statistic.

For the permutation test, each replicate independently draws one of the $4! = 24$ token-order permutations in every cell and recomputes $r$, leaving accuracy fixed. The null is exchangeability of the accuracy–token pairing within cells. Of 100,000 replicates, 2,062 have $|r^{\text{perm}}| \geq |r^{\text{obs}}|$. With the plus-one correction, the two-sided estimate is $p = (2062 + 1)/(100000 + 1) = 0.02063$ (Monte Carlo standard error approximately 0.00045). The observed $r$ is $-0.18067$. The analysis script and extracted run records accompany the source. Calculations use the run accuracies as reported to one decimal place, without replacing the separately reported benchmark means.

Table 10: **Within-cell accuracy–token association.** Each cell fixes family, policy, and benchmark; both metrics are standardized across its four rounds.

| Family | Cells | Runs | Pooled $r$ |
|---|---|---|---|
| DeepSeek | 27 | 108 | $-0.142$ |
| Nemotron | 27 | 108 | $-0.219$ |
| Both | 54 | 216 | $-0.181$ |

Table 11: **Within-cell correlations by benchmark.** Each family contains three policy cells (12 runs) per benchmark; the pooled column contains six cells (24 runs).

| Benchmark | DeepSeek | Nemotron | Pooled |
|---|---|---|---|
| AIME24 | −0.367 | 0.057 | −0.155 |
| AIME25 | −0.494 | −0.194 | −0.344 |
| HMMT25 | −0.086 | −0.151 | −0.119 |
| BRUMO25 | 0.190 | −0.524 | −0.167 |
| CMIMC25 | −0.027 | −0.040 | −0.034 |
| AMC23 | −0.208 | −0.791 | −0.500 |
| MATH | −0.266 | 0.002 | −0.132 |
| Minerva | −0.032 | −0.492 | −0.262 |
| Olympiad | 0.010 | 0.161 | 0.085 |

Eight of nine pooled benchmark correlations are negative; Olympiad is the exception. The family-specific columns also show local sign variation, so the pooled trend is not a universal within-benchmark rule. These are aggregate round-level associations under fixed evaluation settings. They do not show that a correct response is shorter than an incorrect response to the same question, or that imposing a shorter generation causes higher accuracy.

## D CONTROLLER MEASUREMENTS

Table 12: **Full controller sweeps on AIME24.** FC = 0 and 1 denote native donor and anchor reference positions. Tokens are mean reasoning tokens.

| | DeepSeek | | Nemotron | |
|---|---|---|---|---|
| FC / reference | Accuracy (%) | Tokens | Accuracy (%) | Tokens |
| 0.0 (donor) | 47.50 | 7,651 | 60.00 | 9,212 |
| 0.1 | 47.50 | 7,517 | 62.50 | 10,060 |
| 0.2 | 48.33 | 7,639 | 65.00 | 10,420 |
| 0.3 | 49.17 | 7,439 | 66.67 | 10,786 |
| 0.4 | 48.33 | 8,096 | 66.67 | 10,841 |
| 0.5 | 49.17 | 7,795 | 67.50 | 11,159 |
| 0.6 | 50.00 | 7,864 | 65.83 | 11,236 |
| 0.7 | 51.67 | 7,633 | 68.33 | 11,285 |
| 0.8 | **54.17** | 7,798 | **70.00** | 11,395 |
| 0.9 | 50.83 | 7,773 | 68.33 | 11,198 |
| 1.0 (anchor) | 50.83 | 8,065 | 68.33 | 13,030 |

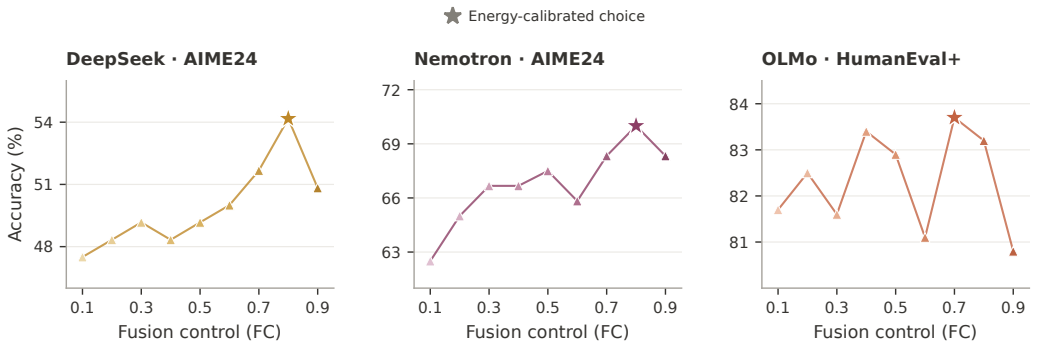



Figure 5: **Accuracy in raw FC coordinates.** Stars denote the energy-calibrated settings. All nine interior settings per family use four evaluation rounds. Native reference positions are omitted from the curves.

Table 13: **Complete HumanEval+ sensitivity measurements.** Every setting uses four evaluation rounds; $\eta$ is measured fusion displacement. MBPP+ is confined to the main-policy comparison.

| FC | Retained energy $\xi_a$ | Displacement $\eta$ | Accuracy (%) | Tokens |
|---|---|---|---|---|
| 0.1 | 0.509 | 0.929 | 81.7 | 4,258 |
| 0.2 | 0.655 | 0.886 | 82.5 | 4,146 |
| 0.3 | 0.754 | 0.846 | 81.6 | 4,338 |
| 0.4 | 0.828 | 0.807 | 83.4 | 4,462 |
| 0.5 | 0.883 | 0.766 | 82.9 | 4,498 |
| 0.6 | 0.924 | 0.724 | 81.1 | 4,591 |
| **0.7** | 0.954 | 0.680 | **83.7** | 4,732 |
| 0.8 | 0.976 | 0.632 | 83.2 | 4,791 |
| 0.9 | 0.991 | 0.582 | 80.8 | 4,939 |

## E QUALITATIVE REASONING EXAMPLES

The four exact-answer cases below compare representative math responses with the source and fused policies side by side. A coding example supplements the HumanEval+ comparison in Table 3. Each correctness count is over four responses per policy. Reasoning-token counts belong to the displayed individual response, not the four-response mean. Cases are selected to make specific success and failure patterns visible, rather than estimate their prevalence. The rounding example illustrates the classification step summarized in the main text.

Table 14: **DeepSeek / MATH #491: counting.** Representative reasoning summaries and four-response correctness counts.

**Problem.** Mr. Potato Head has three hairstyles (or can be bald), two eyebrow pairs, one eye pair, two ear pairs, two lip pairs, and two alternative shoe pairs. All facial parts and one complete shoe pair are required. How many appearances are possible? **Ground truth:** 64.

| **Donor** | **Anchor** |
|---|---|
| Treats mandatory paired parts as optional left/right choices and uses an unsupported product of options. The resulting count does not respect the specified complete pairs. | Recognizes baldness but inconsistently counts the remaining features and shoe alternatives. Its final product undercounts the valid combinations. |
| Answer: 162 (incorrect) | Answer: 32 (incorrect) |
| Selected response: 2,677 reasoning tokens | Selected response: 3,013 reasoning tokens |
| Observed correct: 0/4 | Observed correct: 0/4 |

**SURGE, FC = 0.8**

Includes baldness as a fourth hair choice and treats the two complete shoe pairs as alternatives. It then combines the independent choices without adding optional missing facial parts.

$$4 \cdot 2 \cdot 1 \cdot 2 \cdot 2 \cdot 2 = 64.$$

Answer: 64 (correct) — Observed correct: **3/4**
Selected response: 4,098 reasoning tokens

Table 15: **Nemotron / MATH #119: parentheses.** Representative reasoning summaries and four-response correctness counts.

**Problem.** Insert parentheses in $1+2+3-4+5+6$ to minimize its value, preserving every number, operator, and their order. **Ground truth:** $-9$.

| **Donor** | **Anchor** |
|---|---|
| Introduces a subtraction inside a group containing 5 and 6, changing the original plus sign. Its candidate therefore solves a different expression. | Stops at the example value 3 after considering restricted groupings, while claiming that all possibilities have been checked. It fails to place the entire suffix under the minus sign. |
| Answer: 1 (incorrect) | Answer: 3 (incorrect) |
| Selected response: 6,988 reasoning tokens | Selected response: 10,339 reasoning tokens |
| Observed correct: 1/4 | Observed correct: 1/4 |

**SURGE, FC = 0.8**

Groups all terms after the original minus sign together. This preserves the expression and subtracts the largest possible positive suffix, producing the minimum.

$$(1+2+3)-(4+5+6) = 6-15 = -9.$$

Answer: $-9$ (correct) — Observed correct: **3/4**
Selected response: 7,865 reasoning tokens

Table 16: **Nemotron / Olympiad #544: incenters.** Representative reasoning summaries and four-response correctness counts.

**Problem.** A square $ABCD$ has diagonal 1. Points $E \in AB$ and $F \in BC$ satisfy $\angle BCE = \angle BAF = 30°$, and $G = CE \cap AF$. Find the distance between the incenters of $AGE$ and $CGF$. **Ground truth:** $4 - 2\sqrt{3}$.

| Donor | Anchor |
|---|---|
| Misreads the angle condition as involving the square diagonal, declares a conflict with its $45°$ angle, and then appeals to unsupported symmetry to identify the two incenters. | Uses coordinates and an incenter calculation, but errors in the subsequent coordinate arithmetic lead to an incorrect distance. |
| Answer: $0$ (incorrect) | Answer: $1/2$ (incorrect) |
| Selected response: 10,119 reasoning tokens | Selected response: 17,122 reasoning tokens |
| Observed correct: 0/4 | Observed correct: 0/4 |

**SURGE, FC = 0.8**

Carries the coordinate method through both incenter calculations. With $A = (0, 0)$ and $B = (1/\sqrt{2}, 0)$, the two incenter coordinates differ by $2\sqrt{2} - \sqrt{6}$ in each axis, giving the required distance.

$$\|I_{AGE} - I_{CGF}\| = \sqrt{2}\,(2\sqrt{2} - \sqrt{6}) = 4 - 2\sqrt{3}.$$

Answer: $4 - 2\sqrt{3}$ (correct) Observed correct: **2/4**
Selected response: 19,582 reasoning tokens

Table 17: **Nemotron / Olympiad #480: rounding.** Representative reasoning summaries and four-response correctness counts.

**Problem.** Find the largest real solution of $\lfloor x/3 \rfloor + \lceil 3x \rceil = \sqrt{11}\, x$. **Ground truth:** $189\sqrt{11}/11$.

| Donor | Anchor |
|---|---|
| Tests isolated candidates and extrapolates an unsupported arithmetic pattern. Failure of a later candidate is treated as evidence that the preceding one is maximal. | Sets up the piecewise parametrization, but decimal-division errors underestimate the admissible integer ranges. It stops at a valid, nonmaximal solution. |
| Answer: $29/\sqrt{11}$ (incorrect) | Answer: $84/\sqrt{11}$ (incorrect) |
| Selected response: 14,224 reasoning tokens | Selected response: 17,295 reasoning tokens |
| Observed correct: 0/4 | Observed correct: 0/4 |

**SURGE, FC = 0.8**

Writes $x = 3m + r$ with $0 \le r < 3$ and considers $k = \lceil 3r \rceil$. The final case $k = 9$ permits $m = 18$. Completing this case gives the largest solution; direct substitution yields floor 18 and ceiling 171.

$$x = \frac{10m + k}{\sqrt{11}} = \frac{189}{\sqrt{11}} = \frac{189\sqrt{11}}{11}.$$

Answer: $189\sqrt{11}/11$ (correct) Observed correct: **2/4**
Selected response: 12,238 reasoning tokens

**Why the larger answer is valid.** With $x = 3m+r$ and $k = \lceil 3r \rceil$, the equation becomes $10m+k = \sqrt{11}\, x$. Feasibility gives $m \le 2k$ for $1 \le k \le 8$ and $m \le 18$ for $k = 9$; the latter yields $x = 189/\sqrt{11}$. Its floor and ceiling terms are 18 and 171. Completing this final case, rather than stopping at a smaller feasible candidate, is the decisive mathematical difference.

### E.1 CODING EXAMPLES: SPECIFICATION FIDELITY

The HumanEval+ example in Table 3 and the MBPP+ example below use the energy-calibrated OLMo policy (FC $= 0.7$) and its sources, with four responses per problem. Correctness means passing both the original and additional EvalPlus tests. The tables summarize submitted programs rather than full reasoning traces. They were selected to expose concrete failure modes; they do not estimate how often fusion fixes or introduces errors across the benchmark.

Table 18: **OLMo / MBPP+ #792: counting lists.** SURGE distinguishes the requested type predicate from the shortcut suggested by homogeneous examples.

**Task.** Implement `count_list(lst)` to count top-level elements that are lists. For mixed inputs, counting all elements is incorrect.

| **Donor** | **Anchor** |
|---|---|
| Three responses return the outer length. One uses the correct type test but submits the wrong function name, `count_lists`, and fails the required interface. | Three responses return the outer length, fitting the all-list examples but failing mixed-type inputs. Only one tests whether each element is a list. |
| Observed correct: 0/4 | Observed correct: 1/4 |

**SURGE**

Three responses use `isinstance(item, list)` under the required function name. They count list-valued elements while ignoring scalars. One response still uses the outer-length shortcut.

```
return sum(1 for item in lst if isinstance(item, list))
```

Observed correct: **3/4**

# F SPECTRAL CALIBRATION OF RL UPDATES

The diagnostics below use the same anchor deltas and target matrices as the main fusions: 196 matrices each for DeepSeek and Nemotron, and 224 for OLMo. Nemotron's anchor is step 3440; the energy curve here is measured on that checkpoint. All quantities are computed from weights without generated responses. Equation 6 aggregates squared singular values before taking the ratio, weighting matrices by their update energy.

## F.1 ONE TARGET, DIFFERENT RANK FRACTIONS

Table 19: **Retained anchor-update energy $\xi_a$ across FC.** Bold entries are nearest to $\tau = 0.95$. The Nemotron curve uses step 3440.

| FC | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| DeepSeek | 0.518 | 0.638 | 0.724 | 0.793 | 0.848 | 0.893 | 0.929 | **0.959** | 0.983 |
| Nemotron | 0.405 | 0.546 | 0.652 | 0.736 | 0.805 | 0.862 | 0.909 | **0.947** | 0.978 |
| OLMo | 0.509 | 0.655 | 0.754 | 0.828 | 0.883 | 0.924 | **0.954** | 0.976 | 0.991 |

Applying Equation 7 gives FC $= 0.8/0.8/0.7$ for DeepSeek/Nemotron/OLMo. Their distances from $\tau = 0.95$ are $0.009/0.003/0.004$. The rule picks the nearest candidate, not the first candidate above the target; Nemotron's selected retention is slightly below 0.95. Values are rounded for reporting. No interpolation of the accuracy measurements is used to select FC.

The target was identified from the initial math study and fixed before applying the procedure to OLMo. All three trajectories then use the same construction: compute the anchor RL delta, obtain its retained-energy curve, select the nearest candidate to the shared target, and fuse once. Conditional on the source pair and fixed target, this procedure uses weights alone and requires no accuracy sweep. Its agreement with strong policies across math and code does not establish a universal optimal energy level. The OLMo finer-grid diagnostics, $\xi_a(0.65) = 0.940$ and $\xi_a(0.75) = 0.966$, bracket the selected value; no behavioral measurements are available at those two settings. The four-round sweeps characterize sensitivity (Appendix D).

## F.2 WHY EQUAL RANK FRACTIONS PRESERVE DIFFERENT ENERGY

DeepSeek and Nemotron have 28 layers, with $q/o$ matrices of shape $1536 \times 1536$, $k/v$ of shape $256 \times 1536$, and MLP matrices of shape $1536 \times 8960$ or its transpose. OLMo has 32 layers, square $4096 \times 4096$ attention projections, and $4096 \times 11008$ MLP projections or their transposes. The same retained rank fraction therefore need not capture the same energy even in a shape-matched Gaussian reference.

Table 20: **Actual versus shape-matched reference retention.** Each reference is weighted by the corresponding anchor's module energies.

| | FC $= 0.7$ | | FC $= 0.8$ | |
|---|---|---|---|---|
| Anchor | Actual | Reference | Actual | Reference |
| DeepSeek | .929 | .854 | .959 | .913 |
| Nemotron | .909 | .854 | .947 | .913 |
| OLMo | .954 | .920 | .976 | .958 |

The reference in Table 20 uses same-shape Gaussian matrices and weights their retention by the corresponding anchor's module energy. At FC $= 0.8$, the OLMo reference already retains 95.8%, compared with 91.3% for the math architectures. Its actual update retains 97.6%. Matching $\xi_a$ adjusts for the observed energy distribution rather than assuming that raw FC denotes the same protected content across models. It does not remove every architectural or donor-dependent difference.

### F.3 PROTECTED ENERGY IS DISTINCT FROM FUSION DISPLACEMENT

Retained energy depends only on the anchor. In contrast, the normalized displacement $\eta$ in Equation 8 depends on the donor as well. At the selected settings, $\eta$ is 0.713, 0.559, and 0.680 for DeepSeek, Nemotron, and OLMo. Thus, retaining about 95% of anchor energy does not mean moving 5% of the way to the donor, nor does it predict the donor's behavioral contribution.

The unprotected anchor tails contain 4.1%, 5.3%, and 4.6% of update energy, respectively; 98%, 98%, and 95% of these tails come from MLP projections. With the diagnostic spike threshold set at 1.05 times the same-shape Gaussian spectral edge (scaled to match the median singular value), every selected cutoff falls after the spikes (0/196, 0/196, and 0/224 cutoffs fall within them). These are descriptive spectral diagnostics. A low-energy tail is not necessarily behaviorally irrelevant, and the 95% target is not an intrinsic spectral boundary. In particular, OLMo's residual $q/k$ spectra need not resemble the Gaussian reference.

### F.4 ABSOLUTE-WEIGHT BLOCKS VERSUS UPDATE-DEFINED BLOCKS

Let $\mathcal{K}_{a,k}^{\mathrm{abs},\ell}$ use the leading singular vectors of the absolute anchor weights $W_a^\ell$. Its captured RL-update energy is

$$\zeta_{\mathrm{abs}}(\mathrm{FC}) = \frac{\sum_{\ell\in\mathcal{L}} \|\mathcal{K}_{a,k_\ell}^{\mathrm{abs},\ell}(\Delta_a^\ell)\|_F^2}{\sum_{\ell\in\mathcal{L}} \|\Delta_a^\ell\|_F^2}. \tag{12}$$

For delta-defined blocks this equals $\xi_a(\mathrm{FC})$. Table 21 isolates the comparison for the DeepSeek pair used in the accuracy ablation. Equal block dimensions preserve very different amounts of the RL update, connecting the weight-space diagnostic to the controlled behavioral comparison.

Table 21: **DeepSeek RL-update energy captured at equal block size (%).** The source pair is the one used for the main accuracy ablation.

| FC | Absolute-weight SVD | Delta-SVD |
|---|---|---|
| 0.2 | 7.0 | 63.8 |
| 0.8 | 28.5 | 95.9 |

# G OFFLINE FUSION COST AND TRAINING-STEP COMPARISON

## G.1 MEASURED FUSION TIME

The benchmark uses one RTX A6000 with 48 GB memory, PyTorch 2.6, and FC $= 0.8$. The timed pairs are DeepSeek steps 3600/2600 and Nemotron steps 3440/1200, matching the main math experiments. No 7B latency is inferred from these measurements. A complete pass reads the three checkpoints, computes FP32 delta SVDs and projections for 196 target matrices, evaluates the $\xi$ and $\eta$ diagnostics, and writes BF16 safetensors. Table 22 separates measured warm-cache times from storage and CPU estimates.

Table 22: **One fusion pass on one RTX A6000 (seconds).** Warm-cache entries are measured. The final two rows are estimates, not end-to-end measurements.

| Stage | DeepSeek | Nemotron |
|---|---|---|
| Read three checkpoints (warm page cache) | 0.04 | 0.03 |
| Transfer to GPU | 1.38 | 1.40 |
| 196 FP32 SVDs | 29.45 | 29.56 |
| Projection and fusion | 1.81 | 1.79 |
| Retained-energy curve $\xi$ | 0.07 | 0.07 |
| Displacement curve $\eta$ | 0.34 | 0.34 |
| Write BF16 safetensors | 4.10 | 6.95 |
| **End-to-end, warm cache** | **37.2** | **40.1** |
| Cold-storage estimate (240 MB/s) | 83 | 80 |
| CPU-SVD estimate (48 threads; one-layer extrapolation) | 128 | 124 |

The $\xi$ curve reuses the singular values and takes 0.07 seconds for either family; evaluating the reported $\eta$ curve adds 0.34 seconds. Spectral calibration therefore does not require repeated SVDs or an inference sweep. Given an FC value, only one fused checkpoint is needed. A different FC can reuse the decompositions, although constructing and evaluating additional policies still incurs projection, writing, and inference costs. These timings are individual benchmark measurements, not variability estimates.

## G.2 RECONSTRUCTING ONE RL STEP

We separately measure rollout and actor throughput on the same GPU and reconstruct the resource cost of the JustRL recipe. This is not an end-to-end training run. One step contains $256 \times 8 = 2048$ rollouts, with temperature 1.0 and at most 15,360 response tokens. GRPO uses no critic or KL-reference pass. Actor measurements use FP32 master weights, BF16 autocast, gradient checkpointing, FlashAttention 2.7.4, and micro-batch size one; four mini-batches of 64 prompts yield four optimizer updates. Rollouts use vLLM 0.8.4 and DAPO-Math-17k prompts.

Table 23: **Reconstructed RL-step resource cost (A6000-hours).** Components use single-GPU measurements; distributed overheads are excluded. Totals are as reported; displayed components are rounded.

| Component / reconstruction | DeepSeek | Nemotron |
|---|---|---|
| Rollout, 32-way synchronous accounting | 1.60 | 2.46 |
| Rollout, larger-batch throughput accounting | 0.76 | 1.76 |
| Old-policy log probabilities | 0.14 | 0.32 |
| Actor forward/backward and four AdamW updates | 0.57 | 1.24 |
| Total, synchronous | 2.32 | 4.01 |
| Total, throughput-based | 1.48 | 3.31 |

The synchronous reconstruction multiplies the single-GPU wall time for 64 responses by 32, reflecting the 32-way allocation of a 2048-response step. Those measured times are 180/277 seconds

for DeepSeek/Nemotron. A separate 256-response measurement gives 3,327/3,094 generated tokens per second and mean response lengths 4,472/9,579; mean prompt lengths are 166/174. The throughput-based estimate uses these larger-batch rates instead of the 64-response wall time. Forward and backward token throughput and four AdamW updates supply the actor terms.

Both reconstructions omit weight synchronization, distributed communication, reward scoring, data loading, periodic validation, and training-checkpoint I/O. They are lower-bound accounting estimates under the stated execution assumptions, not measured distributed-training costs. In particular, the synchronous estimate is not a hardware-independent lower bound for asynchronous rollout systems.

**Interpreting the ratio.** The warm-cache fusion resource costs are $37.2/3600 = 0.0103$ and $40.1/3600 = 0.0111$ A6000-hours. They are 0.45%/0.28% of the synchronous reconstruction and 0.70%/0.34% of the throughput-based reconstruction. These compare aggregate GPU resource use, not the wall time of one GPU against 32 GPUs.

**Evaluation is a separate cost.** The timings above cover fusion and weight-space diagnostics, not accuracy evaluation. The included spectral curve supports the weight-only FC rule; evaluating candidate policies is a separate cost. The complete sweeps in this paper characterize sensitivity; they are not mandatory for every checkpoint pair. Deployment can start from a geometrically calibrated candidate and evaluate alternatives as needed on development data. Accuracy evaluation depends on candidate count and decoding protocol; these timings do not measure end-to-end search under the paper's four-response protocol.

## H 7B CODING SETUP AND CHECKPOINT MEASUREMENTS

**Training and initialization.** The RL initialization is Olmo-3-1025-7B. Coding RL follows the OlmoRL recipe on Dolci-RL-Zero-Code-7B, using execution-based verification and a GRPO-derived objective (Ettinger et al., 2025). The recipe uses 32 prompts per batch, eight responses per prompt, temperature 1.0, a constant learning rate of $10^{-6}$, and maximum prompt/response lengths of 2K/16K. It uses active sampling after filtering groups with identical rewards, token-level loss normalization, and asymmetric clipping with lower/upper widths 0.2/0.272. The 3.1 recipe retains truncated sequences in the training loss.

**Evaluation.** HumanEval+ contains 164 problems and MBPP+ contains 378. Decoding uses temperature 1.0, top-$p$ 1.0, and a 32K response limit. Each main policy is evaluated in four rounds with one response per problem per round. The reported accuracy is the mean single-response pass rate, requiring both base and additional EvalPlus tests to pass; it is not success among four attempts. Tokens average the full generated responses. These settings differ from the math decoding configuration, but are fixed across coding sources and fusions.

**Checkpoint roles and coverage.** Table 24 covers 19 measured checkpoints from initialization (step 0) through step 1950. HumanEval+ is evaluated throughout; both coding benchmarks are evaluated for the six later candidates at steps 1300–1700 and 1950. Among these six, step 1950 has the highest two-benchmark mean accuracy and serves as the anchor. Step 1500 supplies a competitive donor: it has the highest native HumanEval+ accuracy and shorter responses on both tasks. These source-policy measurements determine the roles; the fixed spectral target determines FC. The trajectory and sensitivity analyses use HumanEval+ only.

Table 24: **Native coding trajectory on HumanEval+.** Every checkpoint uses four evaluation rounds. SD is the reported standard deviation across round accuracies.

| RL step | Role | Accuracy (%) | SD | Reasoning tokens |
|---|---|---|---|---|
| 0 | Initialization | 42.1 | 3.3 | 2,293 |
| 100 | | 65.1 | 2.3 | 5,090 |
| 200 | | 67.2 | 4.1 | 4,521 |
| 300 | | 70.6 | 3.7 | 4,597 |
| 400 | | 75.0 | 3.6 | 4,609 |
| 500 | | 74.7 | 1.5 | 4,317 |
| 600 | | 76.1 | 2.1 | 4,025 |
| 700 | | 76.2 | 3.1 | 4,320 |
| 800 | | 77.9 | 3.2 | 4,220 |
| 900 | | 79.3 | 1.7 | 4,457 |
| 1000 | | 79.1 | 1.8 | 4,430 |
| 1100 | | 80.0 | 1.5 | 3,892 |
| 1200 | | 80.2 | 2.1 | 4,252 |
| 1300 | | 79.3 | 2.5 | 3,877 |
| 1400 | | 79.9 | 1.0 | 4,372 |
| 1500 | Donor | 82.8 | 1.4 | 4,173 |
| 1600 | | 79.7 | 2.4 | 3,932 |
| 1700 | | 81.6 | 1.0 | 4,271 |
| 1950 | Anchor | 82.0 | 0.4 | 5,275 |

**Reported precision.** Coding accuracies retain the one-decimal precision of the result report. Changes in Table 2 are computed from those displayed means. The 0.0-point MBPP+ change therefore denotes equality at reported precision. The four HumanEval+ round accuracies are 84.1/83.5/83.5/83.5 for SURGE, 82.3/81.7/82.3/81.7 for the anchor, and 84.1/81.1/82.3/83.5 for the donor. All nine FC settings use four rounds (Table 13); no intermediate scores are imputed. Standard deviations retain the source report's rounding.