THE TEACHER IS A DIRECTION, NOT A DESTINATION: EXTRAPOLATING RL-INDUCED REPRESENTATION RESIDUALS IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student to match the teacher's next-token distributions on the student's own trajectories and has yielded substantial empirical gains. Generalized variants allow the student to surpass the teacher by extrapolating an implicit reward in output space. The language-model head, however, attenuates this change anisotropically: much of the change encoded in the teacher's hidden states reaches the logits at a small fraction of its weight, and the sampled-token log-probability ratios on which output-space extrapolation relies inject noise that the extrapolation amplifies, making training unstable. We observe that reinforcement learning (RL) shifts a model's internal representations relative to its base checkpoint, and that the direction of this shift can be measured at every layer. Motivated by this observation, we propose RIDE (**R**L-**I**nduced **D**irection **E**xtrapolation), which extrapolates the RL-induced change directly in representation space: at every layer and token position, RIDE computes the residual between the teacher and its pre-RL checkpoint and regresses the student's hidden states toward targets displaced beyond the teacher along this residual. Conditioned on a sampled trajectory, this regression is equivalent to maximizing a linear directional reward defined by the residual under a quadratic penalty centered at the teacher, which makes explicit how the objective moves the student along the RL-induced direction while limiting its deviation from the teacher. Across four base/RL-teacher pairs spanning different scales, architectures, and pre-training lineages, RIDE approaches or exceeds the RL-trained teacher on every pair and is the only method whose mean does so, and it consistently outperforms output-space extrapolation, which degrades the student whenever the teacher is close to its base.

1 INTRODUCTION

Distillation transfers capability from a teacher to a student (Hinton et al., 2015), and on-policy distillation (OPD) has become a standard stage of language-model post-training (Yang et al., 2025; Xiao et al., 2026). In OPD the student samples its own trajectories and receives per-token supervision from the teacher on the contexts it actually visits, which reduces exposure bias and provides a dense learning signal (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). A common use is to transfer the gains of a reinforcement-learning (RL) teacher to a student that shares its initialization, for example when merging domain experts or distilling an expensive RL run back into its base model (Xiao et al., 2026; Yang et al., 2026b).

Standard OPD treats the teacher as the endpoint of learning, yet students need not stop there: the teacher-to-reference log-probability ratio acts as an implicit reward (Rafailov et al., 2023; Yuan et al., 2025; Cui et al., 2025), up-weighting it relative to the KL constraint yields students that outperform the teacher (Yang et al., 2026b), and related methods extrapolate displacements in parameter or logit space (Li et al., 2026b; Zheng et al., 2025). Because RL updates are small, structured, and consistent across runs (Mukherjee et al., 2025; Zhu et al., 2025; Shenfeld et al., 2026), an RL-trained teacher supplies a direction as well as a destination, and a student that starts from the same checkpoint can continue along it.

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

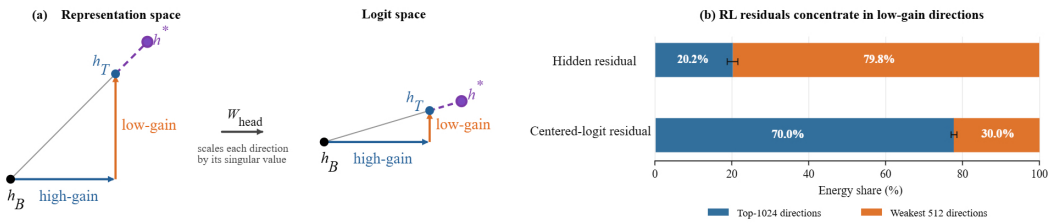


Figure 1: **The language-model head attenuates RL-induced representation changes unevenly.** (a) Residual extrapolation before and after projection through the head. (b) Hidden-state and centered-logit energy shares of the teacher-to-base residual across the singular directions of the head. Whiskers show prompt-bootstrap 95% confidence intervals.

Extrapolating in output space, however, measures this direction after the head has attenuated most of it. The language-model head reweights representation changes according to its anisotropic singular spectrum (Yang et al., 2018; 2026a): in Fig. 1, its 512 weakest directions carry 79.8% of the teacher-to-base residual's hidden-state energy but only 30.0% of its centered-logit energy, so output-space objectives supervise much of the residual only weakly and leave earlier layers unconstrained. The sampled-token log-ratio is moreover length-biased and prone to overoptimization (Park et al., 2024b; Rafailov et al., 2024), and scaling it by a global coefficient amplifies extreme rewards and destabilizes training (Yang et al., 2026b; Sun et al., 2026; Li et al., 2026a). Hidden states retain the change before the head: feature-level distillation already uses them as supervision (Romero et al., 2015; Sun et al., 2019; Jiao et al., 2020), its on-policy form OPRD matches student and teacher states layerwise with a deterministic per-rollout gradient (Yang et al., 2026a), and hidden-state directions encode behaviors that can be steered additively (Park et al., 2024a; Turner et al., 2023; Zou et al., 2023; Arditi et al., 2024), which makes residual extrapolation a natural operation in this space.

We therefore propose *RIDE* (**R**L-**I**nduced **D**irection **E**xtrapolation). With the student initialized at the teacher's pre-RL checkpoint, the three models share a representation space. On each student-generated context, RIDE evaluates the frozen teacher and the pre-RL checkpoint on identical tokens, takes their layerwise hidden-state difference as the RL-induced residual, places each target beyond the teacher along this residual with a single coefficient, and regresses the student's hidden states toward it. Conditioned on a rollout, this regression maximizes a linear directional reward under a quadratic penalty around the teacher, and under a shared linear head the output-space extrapolation target of Yang et al. (2026b) is the head-projected image of the RIDE target; RIDE thus retains the reward-extrapolation view while replacing the sampled-token log-ratio with a deterministic hidden-state gradient.

On four base/RL-teacher pairs (R1-Distill-1.5B, Qwen3-4B, Llama-3.2-3B, and Phi-4-mini), which span different scales, model families, and vocabularies, RIDE attains higher mean Avg@16 on AIME24, AIME25, and AIMO than OPRD and the output-space baselines on every pair and approaches or exceeds the RL-trained teacher itself, whereas output-space extrapolation falls below its teacher on every pair.

Our contributions are as follows.

- We identify the RL-induced representation residual, the layerwise difference between the teacher and its pre-RL checkpoint on identical on-policy contexts, as a measure of how RL changes the model's computation. Both checkpoints are available in the same-initialization distillation setting.
- We propose RIDE, which extrapolates hidden-state targets along this residual with a single coefficient that recovers OPRD at $\lambda = 1$, and we interpret the objective as reward maximization under a quadratic penalty, the representation-space counterpart of output-space reward extrapolation.
- We show that RIDE approaches or exceeds its RL-trained teacher on all four base/teacher pairs while consistently outperforming OPRD and output-space extrapolation, and we characterize its learning signal through analyses of head attenuation and conditional variance.

2 Related Work

On-policy distillation. Unlike sequence-level imitation (Hinton et al., 2015; Kim & Rush, 2016), on-policy distillation samples from the student and supervises the prefixes it visits (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024); its sampled-token form is now a standard post-training stage (Lu & Thinking Machines Lab, 2025; Yang et al., 2025; Xiao et al., 2026). Recent variants remove the need for teacher logits (Ye et al., 2025), reformulate the target for stability (Jang et al., 2026; Jin et al., 2026), analyze failure modes (Fu et al., 2026; Li et al., 2026c), or use the model itself, given privileged information, as its own teacher (Zhao et al., 2026; Hübotter et al., 2026). All of these approaches supervise outputs rather than intermediate states.

Learning beyond the teacher. Viewing OPD as KL-constrained RL with an implicit log-ratio reward (Rafailov et al., 2023; Yuan et al., 2025; Cui et al., 2025), generalized OPD up-weights this reward and obtains students that exceed the teacher (Yang et al., 2026b). A global coefficient, however, drives the student toward extreme reward peaks (Sun et al., 2026) and can collapse near-deterministic outputs (Li et al., 2026a), echoing the length bias and overoptimization of log-ratio rewards (Park et al., 2024b; Rafailov et al., 2024). Other methods extrapolate in parameter or logit space, either along a policy's own RL trajectory (Li et al., 2026b) or along the difference between a fine-tuned model and its initialization (Ilharco et al., 2023; Zheng et al., 2025). RIDE shares the premise that the RL-induced change is a usable direction, but measures it on hidden states, at every layer, before the head, and on the student's own contexts.

Representation-level distillation. Intermediate representations have long served as distillation targets (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Tian et al., 2020; Sun et al., 2019; Jiao et al., 2020; Wang et al., 2020). OPRD applies layerwise hidden-state regression on policy, obtaining a deterministic per-rollout gradient and supervision that output-space losses cannot see through the head (Yang et al., 2026a); PHF instead matches the directions and geometry of hidden-state transitions (Li et al., 2026d). RIDE uses the residual from the pre-RL checkpoint to the teacher to place targets beyond the teacher, drawing on the approximately linear structure of behavioral directions in hidden space (Park et al., 2024a; Turner et al., 2023; Zou et al., 2023; Arditi et al., 2024) and on the structured, consistent nature of RL updates (Mukherjee et al., 2025; Zhu et al., 2025; Shenfeld et al., 2026).

3 Preliminaries

Notation. Let π_θ be the trainable student and π_T the frozen teacher. Both are decoder-only transformers with L layers, hidden dimension d, a shared vocabulary $\mathcal{V}$, and a language-model head $W_{\text{head}} \in \mathbb{R}^{|\mathcal{V}|\times d}$. For a prompt $x \sim \mathcal{D}_x$, the student samples $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T) \sim \pi_\theta(\cdot \mid x)$. We evaluate all models on the same student-generated prefix $(x, \hat{y}_{<t})$ and write $\pi(\cdot \mid x, \hat{y}_{<t})$ for a model's next-token distribution and $h^{(l)}_{\pi,t} \in \mathbb{R}^d$ for its layer-l residual-stream state at position t.

On-policy distillation. Classical knowledge distillation trains the student on teacher outputs or teacher samples (Hinton et al., 2015; Kim & Rush, 2016). On-policy distillation (OPD) instead samples responses from the student and minimizes the reverse KL divergence to the teacher on the prefixes it visits (Agarwal et al., 2024; Gu et al., 2024); by the chain rule of the KL divergence,

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{x,\, \hat{y}\sim\pi_\theta}\Big[\sum_{t=1}^{T} D_{\text{KL}}\big(\pi_\theta(\cdot \mid x, \hat{y}_{<t}) \,\big\|\, \pi_T(\cdot \mid x, \hat{y}_{<t})\big)\Big]. \tag{1}$$

Estimating each local KL from the sampled token gives the commonly used tokenwise update

$$g^{\text{token}}_{\text{OPD}}(\theta) = \mathbb{E}_{x,\, \hat{y}\sim\pi_\theta}\Big[\sum_{t=1}^{T} r_t \,\nabla_\theta \log \pi_\theta(\hat{y}_t \mid x, \hat{y}_{<t})\Big], \qquad r_t = \log \frac{\pi_\theta(\hat{y}_t \mid x, \hat{y}_{<t})}{\pi_T(\hat{y}_t \mid x, \hat{y}_{<t})}, \tag{2}$$

in which $-r_t$ acts as a dense per-token advantage (Lu & Thinking Machines Lab, 2025; Xiao et al., 2026). Relative to any reference policy π_{ref}, the same objective is KL-regularized RL with the implicit reward $\log \pi_T - \log \pi_{\text{ref}}$ (Rafailov et al., 2023; Yang et al., 2026b); output-space extrapolation scales this reward (Section A.1).

On-policy supervision as target matching. Let $s_{\pi,t}$ be a model quantity at position t, τ_t its target, and ℓ a discrepancy measure:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_x,\ \hat{y}\sim\pi_\theta(\cdot|x)}\Big[\frac{1}{T}\sum_{t=1}^{T}\ell\big(s_{\theta,t},\ \mathrm{sg}(\tau_t)\big)\Big], \tag{3}$$

where sg stops gradients through the target. Output-space teacher matching takes $s_{\pi,t}$ to be the next-token distribution and ℓ the reverse KL divergence. Representation matching takes $s_{\pi,t} = h^{(l)}_{\pi,t}$ over a set of layers $l \in \mathcal{L}_{\text{layer}}$ and over positions selected by a mask $m_t \in \{0,1\}$, with the dimension-normalized squared error

$$\ell\big(h^{(l)}_{\theta,t}, \tau^{(l)}_t\big) = \frac{1}{d}\big\|h^{(l)}_{\theta,t} - \tau^{(l)}_t\big\|_2^2, \tag{4}$$

which is on-policy feature distillation (Romero et al., 2015; Jiao et al., 2020; Yang et al., 2026a); conditioned on a rollout, its gradient is deterministic, unlike that of Eq. (2). Both instances set $\tau_t = s_{T,t}$, so the teacher is the endpoint of learning.

Same-initialization setting. We study a teacher obtained by RL from a base checkpoint, $\pi_T = \mathrm{RL}(\pi_B)$, and a student initialized at that checkpoint, $\pi_{\theta_0} = \pi_B$. This setting arises when an RL result is distilled back into its base model or when RL-trained experts that share a base are merged (Xiao et al., 2026; Yang et al., 2026b). The three models share an architecture, a tokenizer, and a representational origin, so on a common prefix the difference $s_{T,t} - s_{B,t}$ measures what RL changed on that context: in output space it is the implicit reward $\log\pi_T - \log\pi_B$ (Rafailov et al., 2023); in representation space it is a vector at every layer, taken before the head.

4 METHOD

RIDE rests on the premise that an RL-trained teacher specifies a direction in addition to a destination: the teacher's displacement from π_B is what RL contributed, and a student initialized at π_B can continue along it (Fig. 2). We state the construction in the target-regression template (Section 4.1), instantiate it on hidden states (Section 4.2), justify this choice of space (Section 4.3), and interpret the objective as reward maximization (Section 4.4; implementation in Section G).

4.1 FROM TEACHER MATCHING TO RESIDUAL EXTRAPOLATION

Within Eq. (3), teacher matching is the choice $\tau_t = s_{T,t}$. We instead displace the target from the teacher along the RL-induced change:

$$\tau_t(\lambda) \;=\; s_{T,t} + (\lambda-1)\big(s_{T,t} - s_{B,t}\big) \;=\; \lambda\, s_{T,t} + (1-\lambda)\, s_{B,t}, \qquad \lambda \geq 0. \tag{5}$$

At $\lambda = 1$ the target is the teacher and Eq. (3) is recovered; for $\lambda \in (0,1)$ it lies between base and teacher; for $\lambda > 1$ it lies beyond the teacher, so the student continues the change that RL initiated. Because $s_{T,t} - s_{B,t}$ is recomputed for every rollout and prefix, the direction is not a fixed offset but the RL-induced change on the contexts the student visits. On log-probabilities, Eq. (5) yields the affine combination $\lambda\log\pi_T + (1-\lambda)\log\pi_B$ that underlies output-space reward extrapolation (Yang et al., 2026b). We argue below that this measures the displacement only after the head has attenuated most of it, and instead instantiate Eq. (5) on hidden states.

4.2 RL-INDUCED REPRESENTATION RESIDUAL AND THE RIDE OBJECTIVE

Given a student rollout, the frozen models π_T and π_B are evaluated on the same prefix $(x, \hat{y}_{<t})$, and the *RL-induced residual* at layer l and position t is $\Delta^{(l)}_t = h^{(l)}_{T,t} - h^{(l)}_{B,t}$. Because both models process the same tokens, $\Delta^{(l)}_t$ isolates what RL changed in the processing of that context rather than the effect of a different trajectory; it lives in the student's hidden-state space, is defined at every layer, and is computed deterministically from two forward passes rather than sampled.

Applying Eq. (5) to hidden states yields the RIDE target,

$$h^{\star(l)}_t \;=\; h^{(l)}_{T,t} + (\lambda-1)\,\Delta^{(l)}_t \;=\; \lambda\, h^{(l)}_{T,t} + (1-\lambda)\, h^{(l)}_{B,t}, \tag{6}$$

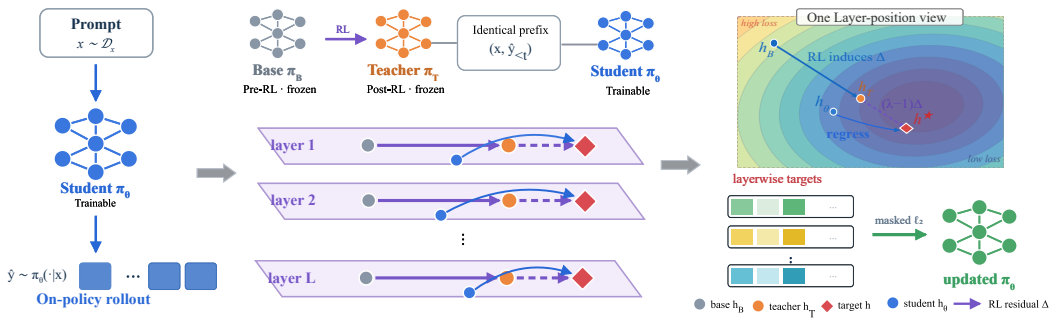


Figure 2: **Overview of RIDE.** On student-generated prefixes, the frozen base and teacher define layerwise residuals. RIDE extrapolates hidden-state targets along these residuals and trains the student by masked regression toward them. Right: target geometry at a single layer and position.

and the training objective is the representation-space regression of Eq. (4) toward this target,

$$\mathcal{L}_\lambda(\theta) = \mathbb{E}_{x,\, \hat{y}\sim\pi_\theta}\Big[\frac{1}{|\mathcal{L}_{\text{layer}}|}\sum_{l\in\mathcal{L}_{\text{layer}}}\frac{1}{M}\sum_{t=1}^{T} m_t\,\frac{1}{d}\big\|h_{\theta,t}^{(l)} - \text{sg}\big(h_t^{\star(l)}\big)\big\|_2^2\Big], \tag{7}$$

where $M = \sum_t m_t$. Unlike its log-probability counterpart, Eq. (6) involves no partition function. We supervise all L layers and the final k response positions, where student–teacher disagreement concentrates (Yang et al., 2026a). Only the target differs from representation matching, so $\lambda = 1$ recovers OPRD exactly and any difference between the two methods is due solely to the residual term. Because Eq. (4) is unnormalized, the loss scale grows with λ; we keep this form so that $\lambda = 1$ reduces exactly to OPRD and rescale the loss coefficient instead (Remark G.1).

4.3 Why the Displacement Should Be Measured in Hidden States

Two arguments favor hidden states over log-probabilities in Eq. (5): how much of the residual survives the head, and how estimation noise behaves when the target is displaced beyond the teacher.

The head attenuates the residual anisotropically. Let $\tilde{h}$ denote the representation fed to the head and $z = W_{\text{head}}\tilde{h}$ the logits. Since the head is linear, displacing its input along the residual displaces the logits along the corresponding logit residual,

$$W_{\text{head}}\big(\tilde{h}_T + (\lambda-1)\tilde{\Delta}\big) = z_T + (\lambda-1)(z_T - z_B) = \lambda z_T + (1-\lambda)z_B. \tag{8}$$

Up to the log-partition constant removed by the softmax, the right-hand side is the log-probability target of Eq. (5); the output-space target is thus the image of the RIDE target under the head. The converse fails for two reasons. First, a logit target constrains $\tilde{h}$ only weakly along the directions that W_{head} amplifies least, and Section 5.4 shows that the RL-induced residual concentrates in precisely these directions, so they reach an output-space loss at a small fraction of their weight. Second, a logit target imposes no constraint on layers below the final one, whereas Eq. (6) specifies the full vector at every supervised layer. Output-space reward extrapolation is therefore the head projection of representation extrapolation, which extrapolates the residual at full weight in every direction; the assumptions behind Eq. (8) and the head-drift term are discussed in Section E.

Extrapolation amplifies output-space noise but not hidden-state noise. Fix a rollout and a position t; write $p = \pi_\theta(\cdot \mid x, \hat{y}_{<t})$, $q = \pi_T(\cdot \mid x, \hat{y}_{<t})$, and $b = \pi_B(\cdot \mid x, \hat{y}_{<t})$, and let $v \sim p$ be the sampled token. With $\ell(v) = \log p(v) - \log q(v)$ and $\rho(v) = \log q(v) - \log b(v)$, output-space extrapolation uses the advantage $A_\lambda(v) = \ell(v) - (\lambda-1)\rho(v)$ in the tokenwise update of Eq. (2) (Section A.1); both components are evaluated at a single sampled token.

Proposition 4.1 (Conditional variance under extrapolation)**.** *Conditioned on the prefix, the variance of the output-space advantage over the sampled token satisfies*

$$\text{Var}_{v\sim p}\big[A_\lambda(v)\big] = \text{Var}_{v\sim p}[\ell(v)] + (\lambda-1)^2\,\text{Var}_{v\sim p}[\rho(v)] - 2(\lambda-1)\,\text{Cov}_{v\sim p}[\ell(v), \rho(v)]. \tag{9}$$

The second term does not vanish as $p \to q$ unless $\log q - \log b$ is constant on the support of q. Conditioned on the same prefix, the gradient of the RIDE objective in Eq. (7) is deterministic for every λ, so its conditional variance is zero.

The proof is in Section A. In output space the extrapolated component is a sampled estimate of ρ whose noise is scaled by $(\lambda-1)^2$ however close the student is to the teacher, which accounts for the instability reported at large scaling factors (Yang et al., 2026b; Sun et al., 2026; Li et al., 2026a); in hidden-state space the residual is computed rather than sampled, so λ changes only the destination.

4.4 Optimization Interpretation

The regression in Eq. (7) can be read as reward maximization under a proximity constraint. Fix a rollout, a supervised layer, and a position, and write $h = h^{(l)}_{\theta,t}$, $h_T = h^{(l)}_{T,t}$, and $\Delta = \Delta^{(l)}_t$.

Proposition 4.2 (Reward-and-penalty form)**.** *Conditioned on the rollout, minimizing $\mathcal{L}_\lambda$ with respect to the student's hidden state at each supervised layer and position is equivalent, up to a positive multiplicative factor and an additive constant, to maximizing*

$$(\lambda-1)\, r(h) \;-\; \tfrac{1}{2}\big\|h - h_T\big\|_2^2, \tag{10}$$

where $r(h) = \langle h - h_T, \Delta\rangle$. The maximizer is $h^\star = h_T + (\lambda-1)\Delta$, and $\nabla_h \mathcal{L}_\lambda = \frac{2}{d}(h - h^\star)$.

The reward r is the inner product of the student's displacement from the teacher with the RL-induced residual, the penalty is quadratic in the same displacement, and $\lambda - 1$ weights one against the other; at $\lambda = 1$ the reward weight vanishes and OPRD is recovered. This is the representation-space counterpart of the KL-constrained RL view of output-space extrapolation (Yang et al., 2026b); see Section A.3 for the derivation.

5 Experiments

5.1 Experimental Setup

Models. We use four base/RL-teacher pairs; in each, the student is initialized from the base checkpoint (Section 3). The R1-Distill-1.5B pair consists of DeepSeek-R1-Distill-Qwen-1.5B and JustRL-DeepSeek-1.5B (Guo et al., 2025; He et al., 2025). The remaining teachers, Just-Qwen3-4B, Just-Llama-3.2-3B, and Just-Phi-4-mini, are obtained by applying the same RL recipe (He et al., 2025) to Qwen3-4B (Yang et al., 2025), Llama-3.2-3B (Grattafiori et al., 2024), and Phi-4-mini (Abouelenin et al., 2025). Within each pair, the base, teacher, and student share the architecture, tokenizer, and language-model head, which permits direct hidden-state comparison; across pairs, depth, width, vocabulary, and pre-training lineage all differ.

Training data and protocol. Prompts are drawn from DAPO-Math-17K (Yu et al., 2025). Each method samples rollouts from its own current student policy, and the frozen teacher and, where needed, the base checkpoint are evaluated on those prefixes. All runs share the prompt dataset, sampling settings, optimizer schedule, and a budget of 500 training steps. Unless stated otherwise, Sections 5.3 to 5.5 use the R1-Distill-1.5B pair, and RIDE supervises all L layers and the last $k = 2{,}000$ response positions with $\lambda = 1.25$ and loss scaling by λ^{-2} (Remark G.1). Hyperparameters and hardware are listed in Section G.

Baselines. We compare against sampled-token OPD (top-1) and top-16 OPD for output-space teacher matching (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025); OPRD for representation-space matching, which coincides with RIDE at $\lambda = 1$ (Yang et al., 2026a); and ExOPD for output-space extrapolation, which uses the same λ and pre-RL checkpoint (Yang et al., 2026b). The frozen teacher and the untouched student serve as reference points. Scaling coefficients are swept over the same grid as for RIDE. We omit weight-space extrapolation because ExOPD was stronger in the same setting (Zheng et al., 2025; Yang et al., 2026b).

Evaluation. We report Avg@16 on AIME 2024, AIME 2025, and AIMO (AMC 2022–2023) (AI-MO, 2024a; OpenCompass, 2025; AI-MO, 2024b). For each training run, we average correctness

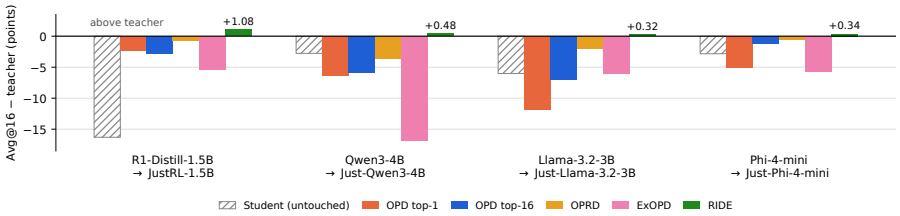


Figure 3: Avg@16 relative to the RL-trained teacher (points; the horizontal line is the teacher) for every method and base/teacher pair in Table 1. RIDE is the only method whose mean lies above the line on all four pairs; across-seed standard deviations are in Table 4.

over 16 sampled responses per problem and then over problems within each benchmark; Avg. is the unweighted mean of the three benchmark scores. We use three independent training seeds (14, 42, and 2027) for each trained method in Table 1, which reports the mean of the run-level scores. The teacher and untouched student are fixed-checkpoint reference evaluations, so their variation is not training-seed variation. Sampling and grading details are in Section G.

5.2 Main Results

Table 1 reports Avg@16 on the four base/RL-teacher pairs, Fig. 3 plots each method's average relative to its teacher, and Table 4 reports the across-seed standard deviations for OPRD and RIDE. RIDE is the only method whose mean lies above the teacher line, on every pair, although on three pairs the margin is within one across-seed standard deviation (Table 4); all teacher-matching methods remain below the line, as their objectives dictate, and output-space extrapolation falls below it as well. Since RIDE differs from OPRD only in its target, its margin over OPRD (0.97 to 4.06 points) isolates the benefit of residual extrapolation under a shared protocol.

Figure 3 also separates the space in which extrapolation is performed from extrapolation itself: ExOPD and RIDE use the same λ and the same pre-RL reference and differ only in where the displacement is measured. ExOPD falls below its teacher on every pair, and on the R1-Distill pair it also falls below the output-space teacher-matching baselines (49.87 versus 52.90 for sampled-token OPD) even though RL moved this teacher 16.3 points from its base; on the three pairs where that change is small (2.8 to 6.0 points) it falls below the untouched student as well, by 14.1 points on Qwen3-4B. This is the behavior predicted by Proposition 4.1: the extrapolated component of the sampled advantage is scaled by $(\lambda-1)^2$ regardless of how much signal the log-ratio carries, and the damage is largest when the teacher-to-base gap is small. RIDE instead improves over the student and over OPRD on every pair, and over ExOPD by 9.1 points on average. On the Llama-3.2-3B pair the AIME benchmarks lie near the evaluation floor, so the evidence there rests mainly on AIMO.

5.3 Effect of the Extrapolation Coefficient

Figure 4 compares RIDE and ExOPD on the R1-Distill-1.5B pair for $\lambda \in \{0.5, 0.75, 1, 1.25, 1.5, 2\}$; at $\lambda = 1$ they reduce to OPRD and to sampled-token OPD, so each row isolates extrapolation within one space. For RIDE, final Avg@16 rises with λ from 48.2 at $\lambda = 0.5$ to 54.3 at 1 and peaks at 55.4 at 1.25; the neighboring coefficients $\lambda = 1.15$ and 1.35 give 55.2 and 55.3, so every $\lambda \in [1.15, 1.35]$ improves over OPRD (full numbers in Section B). Every run with $0.75 \le \lambda \le 1.35$ improves steadily over the whole budget with a format score of 94–97%, and beyond $\lambda = 1.35$ the degradation is graceful, with $\lambda = 1.5$ and 2 finishing at 52.3 and 52.4. Displacing beyond the teacher ($\lambda \ge 1$) shortens responses to 5.3–5.7k tokens, whereas displacing toward the base ($\lambda < 1$) lengthens them to 6.8–7.4k. ExOPD is best at $\lambda \le 1$ (52.9 at 0.75 and at 1) and is harmed by every $\lambda > 1$: the extrapolated runs reach their peak early (by step 150 for $\lambda = 1.25$ and 1.5) and then decline ($\lambda = 1.25$ to 49.9, $\lambda = 2$ to 46.2), the format score falls from 92% at $\lambda = 0.5$ to 64% at 2, and responses lengthen to 8,700 tokens. This is the behavior predicted by Proposition 4.1: the sampled-advantage noise scales with $(\lambda-1)^2$, so the coefficient that moves RIDE further along the RL-induced direction moves ExOPD off the teacher's formatting and below its teacher-matching baseline. Elsewhere we use $\lambda = 1.25$ for both methods, the best value for RIDE and the smallest extrapolating value for ExOPD.

Table 1: Main results (Avg@16, %) on four base/RL-teacher pairs. Trained-method entries are means over three independent training seeds; the teacher and untouched student are fixed-checkpoint reference evaluations. Each run uses its own on-policy rollouts; within a pair, all runs share prompts, sampling settings, and optimizer schedule. ExOPD and RIDE use $\lambda = 1.25$ with the pre-RL checkpoint as reference. Bold marks the best distillation method per column.

	R1-Distill-1.5B → JustRL-1.5B				Qwen3-4B → Just-Qwen3-4B			
Method	AIME24	AIME25	AIMO	Avg.	AIME24	AIME25	AIMO	Avg.
Teacher	50.80	35.60	79.50	55.30	63.13	51.04	82.61	65.59
Student	32.90	21.90	62.20	39.00	61.25	48.75	78.46	62.82
OPD top-1	47.10	33.50	78.10	52.90	52.50	45.83	79.40	59.24
OPD top-16	47.10	34.00	76.50	52.53	52.92	47.08	79.22	59.74
OPRD	49.80	34.60	79.10	54.50	61.46	44.38	80.20	62.01
ExOPD	44.20	29.60	75.80	49.87	41.88	28.54	75.83	48.75
RIDE	**50.21**	**37.29**	**81.63**	**56.38**	**63.13**	**52.92**	**82.15**	**66.07**
	Llama-3.2-3B → Just-Llama-3.2-3B				Phi-4-mini → Just-Phi-4-mini			
Method	AIME24	AIME25	AIMO	Avg.	AIME24	AIME25	AIMO	Avg.
Teacher	13.75	0.42	24.85	13.01	11.46	6.04	36.37	17.96
Student	4.17	0.63	16.19	6.99	7.92	4.79	32.68	15.13
OPD top-1	0.63	0.21	2.41	1.08	6.25	3.75	28.69	12.90
OPD top-16	2.08	0.63	15.21	5.97	10.00	5.63	34.64	16.76
OPRD	7.92	1.04	23.87	10.94	**10.21**	6.46	35.32	17.33
ExOPD	3.33	0.42	17.09	6.95	7.29	1.67	27.71	12.22
RIDE	**13.50**	**1.30**	**25.20**	**13.33**	**10.21**	**7.50**	**37.20**	**18.30**

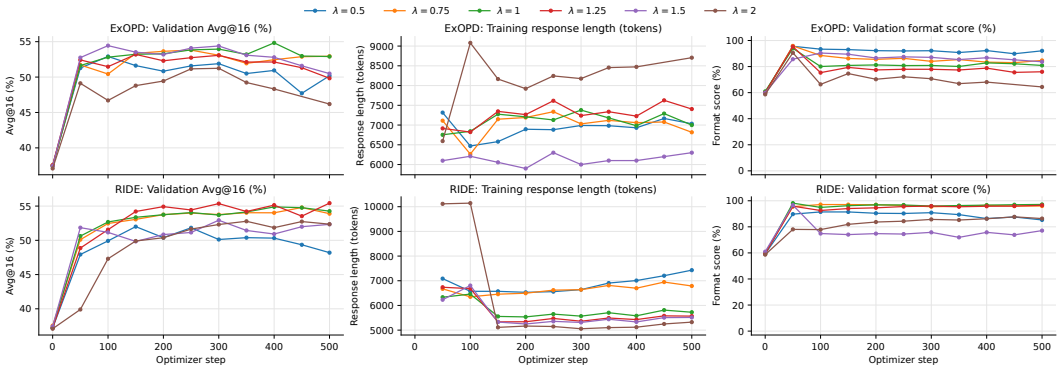
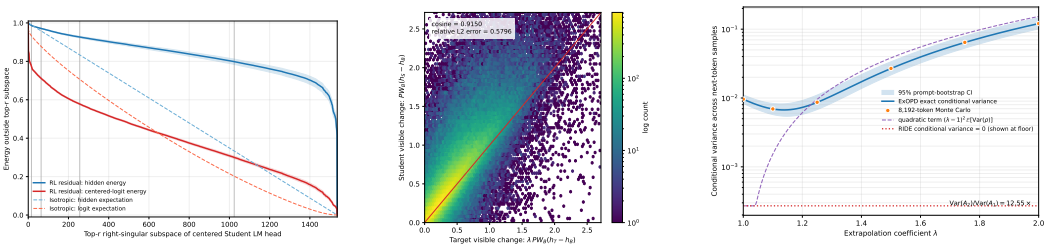


Figure 4: ExOPD (top) and RIDE (bottom) at six extrapolation coefficients on the R1-Distill-1.5B pair (one run each): validation Avg@16 (%) averaged over AIME24, AIME25, and AIMO; mean response length over the preceding 50 steps; validation format score.

5.4 Mechanistic Analysis

Head attenuation reaches the learning signal. The residual retains only 0.59 of the head gain of an equal-norm isotropic direction, and its energy concentrates in the weakest head directions (Fig. 5a); these directions receive 73.0% of the representation-loss gradient but only 48.6% of the output-KL gradient at the head input, so head attenuation weakens the output-space learning signal as well. The centered head is full rank, with singular values spanning 71.7 to 1.8, so the residual is attenuated by a factor of up to 40 along its dominant directions rather than removed: an output-space objective still sees the RL-induced change, but at a fraction of its weight. The student's update aligns with the residual (cosine 0.954) and its head-visible component tracks the target (Fig. 5b, cosine 0.895), with a projection onto the residual of 1.69, above the target coefficient of 1.25: λ specifies the target rather than the realized displacement, and the student continues past the target along the same direction. During RL the teacher head itself drifts from W_B by only 1.65%, which supports the shared-head assumption behind Eq. (8); the representation-space target contains no head-drift term. Further measurements are in Section E.



(a) Residual energy outside top-r head directions.

(b) Student versus target change.

(c) Conditional variance across λ.

Figure 5: Mechanistic analysis on the R1-Distill-1.5B pair ($\lambda = 1.25$, step 500, final layer). (a) The 512 weakest head directions hold 79.8% of the residual's hidden-state energy versus 33.3% under isotropy. (b) Realized student change versus target. (c) Next-token conditional variance of the ExOPD advantage and the RIDE gradient norm across λ.

The learning signal is deterministic in representation space. At a fixed prefix, the sampled-token advantage of ExOPD varies with the next token, whereas the RIDE gradient does not (Proposition 4.1). On 9,152 on-policy validation prefixes per method (Fig. 5c), the conditional variance of ExOPD rises $12.6\times$ from $\lambda = 1$ to $\lambda = 2$, with the extrapolation term dominating beyond $\lambda \approx 1.3$, whereas resampling the next token leaves the RIDE per-position gradient norm unchanged at every prefix and every λ (drawn at the floor of the logarithmic axis). The decomposition of Eq. (9) also explains why ExOPD tolerates coefficients slightly above one: the covariance term is negative, so the variance has a shallow minimum near $\lambda \approx 1.14$ before the $(\lambda-1)^2 \operatorname{Var}[\rho]$ term takes over, which matches the sweep in Section 5.3, where ExOPD degrades at $\lambda = 1.25$ and collapses at $\lambda = 2$. The sampling protocol, variance decomposition, and uncertainty estimates are in Section F.

5.5 Ablation: The Residual Direction

A displaced target could help for reasons unrelated to the RL-induced residual, for instance by acting as a regularizer or by enlarging the loss scale. We therefore ablate the residual direction on the R1-Distill-1.5B pair with four controls that hold the displacement magnitude fixed and alter only its direction (Table 3 in Section C). The *random* control replaces $\Delta_t^{(l)}$ with a Gaussian vector of the same norm; the *reversed* control uses $\lambda < 1$, displacing the target toward π_B; the *mismatched-origin* control computes the residual against Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024), which shares the architecture but not the training lineage, so that $h_T - h_{B'}$ reflects a difference between unrelated models rather than the change produced by RL; and the *trajectory-mismatched* control computes h_T and h_B on separately generated trajectories rather than on the identical student prefix, testing the claim in Section 4.2 that the fixed context gives the residual its meaning. Relative to OPRD (54.50), the random, reversed, mismatched-origin, and trajectory-mismatched controls reach 55.04, 53.90, 54.75, and 55.12, all within one point, whereas RIDE with the RL-induced residual reaches 56.38.

6 Conclusion

An RL-trained teacher supplies a direction as well as a destination. The layerwise hidden-state difference between the teacher and its pre-RL checkpoint on identical prefixes measures what RL changed, and RIDE displaces the representation-matching target beyond the teacher along it; only the target changes, so $\lambda = 1$ recovers OPRD. Measuring it before the head matters because the head attenuates the residual anisotropically and leaves earlier layers unconstrained, and because output-space extrapolation amplifies sampled-token noise by $(\lambda - 1)^2$ whereas the RIDE gradient is deterministic given the rollout. Across four base/teacher pairs, RIDE approaches or exceeds its teacher, consistently outperforms OPRD and output-space extrapolation, remains stable where the latter collapses, and, as the direction controls show, owes its gain to the RL-induced residual itself. RIDE requires the pre-RL checkpoint and a shared representation space, uses a single global λ, and was evaluated only on mathematical reasoning with one RL recipe. Relaxing these constraints and studying the safety and calibration of students trained on hidden-state targets are left for future work.

Ethics statement

This work studies a training objective for distilling mathematical-reasoning ability from an RL-trained teacher into a student that shares its initialization. All experiments use publicly released model checkpoints and public prompt and evaluation sets (DAPO-Math-17K, AIME 2024, AIME 2025, and the AIMO AMC 2022–2023 problems); no human subjects, personal data, or proprietary data are involved, and the task domain (competition mathematics) carries no direct risk of harm. Two considerations nonetheless apply. First, like any distillation method, RIDE transfers the teacher's behavior, and therefore also any biases or errors that RL introduced into it, to the student; because RIDE displaces the target *beyond* the teacher, such properties could in principle be amplified rather than merely copied. Second, the student is supervised on hidden-state targets that no model has actually produced, so its calibration and safety behavior are not guaranteed to match those of the teacher and should be evaluated independently before deployment; we note this limitation in Section 6. We do not foresee misuse specific to this method beyond that of on-policy distillation in general, and we report the compute used (Table 5) so that the environmental cost of reproducing our results can be assessed.

Reproducibility statement

We have taken the following steps to make our results reproducible. The RIDE objective is fully specified by Eqs. (6) and (7) together with the loss rescaling in Remark G.1, and its implementation on top of an OPRD pipeline is described in Section G, including how the pre-RL checkpoint is hosted and how the target is formed and transmitted. Table 5 lists every training and evaluation hyperparameter shared by all methods (prompt set, rollout settings, optimizer, schedule, precision, hardware, supervised layers and positions, λ, evaluation temperature, and benchmark sizes), and Table 6 states the per-rollout cost of each method. All base checkpoints and the JustRL-DeepSeek-1.5B teacher are publicly available; the remaining teachers are obtained by applying the published JustRL recipe (He et al., 2025) to the corresponding public base models, as described in Section 5.1. Complete proofs of Propositions 4.1 and 4.2 and the derivation of the extrapolated advantage are given in Section A. The full extrapolation-coefficient sweep behind Fig. 4, including per-benchmark numbers, peak checkpoints, format scores, and response lengths, is tabulated in Table 2, and the sampling protocols and sample sizes for every diagnostic in Section 5.4 are specified in Sections E and F. Training code, the RL-trained teacher checkpoints, and evaluation scripts will be released upon publication.

AI use statement

In this work, we used generative AI tools to assist with language polishing, checking consistency of terminology and notation, reviewing the presentation and interpretation of reported results, and verifying bibliographic metadata against source records. The tools did not design or run experiments or generate the reported measurements, figures, or tables. The authors are responsible for the research decisions, technical claims, and final manuscript, including AI-assisted edits.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

AI-MO. AIMO validation AIME. `https://huggingface.co/datasets/AI-MO/aimo-validation-aime`, 2024a.

AI-MO. AIMO validation AMC. `https://huggingface.co/datasets/AI-MO/aimo-validation-amc`, 2024b.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322.

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

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, et al. JustRL: Scaling a 1.5B LLM with a simple RL recipe. *arXiv preprint arXiv:2512.16649*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *International Conference on Machine Learning*, 2026.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227, 2026. doi: 10.18653/v1/2026.findings-acl.2094.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, 2020.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *International Conference on Machine Learning*, 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *International Conference on Machine Learning*, volume 235, pp. 24872–24895, 2024.

Xin Li, Hao Jiang, Annan Wang, Yichi Zhang, and Chau Yuen. The extrapolation cliff in on-policy distillation of near-deterministic structured outputs. *arXiv preprint arXiv:2605.08737*, 2026a.

Yang Li, Semih Yavuz, and Shafiq Joty. RISE: Recursive improvement via self-extrapolating policy distillation. *arXiv preprint arXiv:2609.05295*, 2026b.

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

Yuhan Li, Mingxu Zhang, Dazhong Shen, and Ying Sun. PHF: Privileged hidden flow for on-policy self-distillation. *arXiv preprint arXiv:2606.29340*, 2026d.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tür, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 132119–132138, 2025. doi: 10.52202/085713-4399.

OpenCompass. AIME 2025. https://huggingface.co/datasets/opencompass/AIME2025, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, volume 235, pp. 39643–39666, 2024a. URL https://proceedings.mlr.press/v235/park24c.html.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126207–126242, 2024. doi: 10.52202/079017-4009.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations*, 2026.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4323–4332, 2019.

Yang Sun, Lichao Ma, Houyuan Qin, Yuxin Liu, Hanyang Lu, Yao Zhu, Pinlong Cai, and Guohang Yan. REOPD: Reliability-adaptive reward extrapolation for on-policy distillation. *arXiv preprint arXiv:2608.11698*, 2026.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. MiMo-V2-Flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. OPRD: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026a.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244, 2025. doi: 10.52202/085713-3775.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *International Conference on Machine Learning*, volume 267, pp. 73511–73525, 2025. URL `https://proceedings.mlr.press/v267/yuan25c.html`.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *International Conference on Machine Learning*, 2026.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. Model extrapolation expedites alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1025–1041, 2025. doi: 10.18653/v1/2025.acl-long.51.

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

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

A Derivations and Proofs

A.1 On-Policy Distillation as Implicit-Reward Reinforcement Learning

Fix a prefix and write $p = \pi_\theta(\cdot \mid x, \hat{y}_{<t})$, $q = \pi_T(\cdot \mid x, \hat{y}_{<t})$, and let π_{ref} be any reference policy with full support. Adding and subtracting $\log \pi_{\text{ref}}$ inside the local KL of Eq. (1) gives

$$D_{\text{KL}}(p \,\|\, q) = D_{\text{KL}}(p \,\|\, \pi_{\text{ref}}) - \mathbb{E}_{v\sim p}\big[\log q(v) - \log \pi_{\text{ref}}(v)\big], \tag{11}$$

so minimizing the OPD loss is KL-regularized RL with the implicit reward $\log q - \log \pi_{\text{ref}}$, in which the reward and the KL term have equal weight (Yang et al., 2026b). Weighting the reward by λ instead, the local objective $\lambda\, \mathbb{E}_{v\sim p}[\log q - \log \pi_{\text{ref}}] - D_{\text{KL}}(p \,\|\, \pi_{\text{ref}})$ is maximized by

$$\log p^\star = \lambda \log q + (1-\lambda) \log \pi_{\text{ref}} - \log Z, \tag{12}$$

where Z normalizes $p^\star$. With $\pi_{\text{ref}} = \pi_B$ this is Eq. (5) on log-probabilities. Minimizing $D_{\text{KL}}(p \,\|\, p^\star)$ with the sampled-token estimator of Eq. (2) uses the per-token quantity

$$\log p(v) - \log p^\star(v) = \ell(v) - (\lambda - 1)\rho(v) + \log Z, \tag{13}$$

with ℓ and ρ as in Section 4.3. The constant $\log Z$ does not depend on v and multiplies a score function with zero conditional mean, so the extrapolated advantage is $A_\lambda(v) = \ell(v) - (\lambda-1)\rho(v)$. The tokenwise update treats sampled prefixes as fixed; differentiating the sequence-level objective in Eq. (1) would additionally account for how earlier actions change later prefixes.

A.2 Proof of Proposition 4.1

For random variables X, Y and a constant c, $\text{Var}[X - cY] = \text{Var}[X] + c^2\text{Var}[Y] - 2c\,\text{Cov}[X, Y]$. Taking $X = \ell(v)$, $Y = \rho(v)$, $c = \lambda - 1$, and $v \sim p$ gives Eq. (9). Since q and b are frozen, ρ does not depend on θ, and the second term grows quadratically in $\lambda - 1$. As $p \to q$, $\text{Var}_{v\sim p}[\rho(v)] \to \text{Var}_{v\sim q}[\rho(v)]$, which vanishes only if $\log q - \log b$ is constant on the support of q, that is, only if RL left the next-token distribution at this prefix unchanged.

For RIDE, the student state $h^{(l)}_{\theta,t}$ is a deterministic function of θ and the prefix, and the target $h^{\star(l)}_t$ is a deterministic function of the prefix and the two frozen models. Neither depends on the token sampled at position t, so the per-position loss and its gradient with respect to θ are invariant to resampling that token, and their conditional variance is zero.

A.3 Reward-and-Penalty Form: Derivation of Proposition 4.2

With h, h_T, and Δ as in Section 4.4, expanding the squared distance to the target gives

$$\big\|h - h_T - (\lambda-1)\Delta\big\|_2^2 = \big\|h - h_T\big\|_2^2 - 2(\lambda-1)\big\langle h - h_T,\ \Delta\big\rangle + (\lambda-1)^2\|\Delta\|_2^2, \tag{14}$$

in which the final term is independent of θ. The squared distance therefore equals $\|h - h_T\|_2^2 - 2(\lambda-1)\, r(h)$ plus a constant, and multiplying by $-\frac{1}{2}$ gives Eq. (10) up to an additive constant; the per-position loss is this squared distance scaled by the positive factor $1/d$. The objective in Eq. (10) is strictly concave in h with gradient $(\lambda-1)\Delta - (h - h_T)$, which vanishes at $h^\star = h_T + (\lambda-1)\Delta$. Differentiating $\frac{1}{d}\|h - h^\star\|_2^2$ gives $\frac{2}{d}(h - h^\star)$.

The reward r measures the extent to which the student's displacement from the teacher is aligned with the RL-induced residual, the penalty is quadratic in the same displacement, and $\lambda - 1$ weights the reward against the penalty. This mirrors the view of output-space extrapolation as KL-constrained RL around the teacher (Yang et al., 2026b), with the KL constraint replaced by a quadratic one, the log-ratio reward by an inner product with the residual, and the score-function estimator by a vector gradient. At $\lambda = 1$ the reward weight vanishes: teacher matching is a special case of RIDE rather than its point of departure.

B Extended Extrapolation-Coefficient Sweep

Table 2 lists the numbers behind Fig. 4: validation Avg@16 on AIME24, AIME25, and AIMO at the final checkpoint, the validation format score, and the mean training response length over the 50

steps preceding the final checkpoint. The two additional RIDE coefficients $\lambda = 1.15$ and 1.35 are included; they bracket $\lambda = 1.25$ and both remain above OPRD (55.2 and 55.3 versus 54.3), so every coefficient in $[1.15, 1.35]$ improves over teacher matching, with $\lambda = 1.25$ the best. The sweep uses a single training seed per coefficient, whereas Table 1 averages three seeds; the $\lambda = 1$ and $\lambda = 1.25$ rows of Table 2 are therefore single runs (54.3 and 55.4) and differ from the corresponding three-seed means in Table 1 (54.50 and 56.38) by about one across-seed standard deviation (Table 4). Figure 6 shows the final Avg@16 on each benchmark as a function of λ: the gain of RIDE around $\lambda = 1.25$ and the decline of ExOPD beyond $\lambda = 1$ appear on all three benchmarks, not only in the macro average. Figure 7a places every run in the length–accuracy plane; increasing λ moves RIDE toward shorter and more accurate responses with a high format score, whereas it moves ExOPD toward longer and less accurate ones. Figure 7b summarizes the last 200 training steps of each run in Fig. 4 by the mean and the min–max range of validation Avg@16 over steps 300–500. For RIDE, the late-training mean rises from 54.3 at $\lambda = 1$ to 54.7 at $\lambda = 1.25$, and every run with $\lambda \geq 0.75$ stays above 50.9 throughout the window; for ExOPD, the late-training mean of every $\lambda > 1$ run lies below that of $\lambda = 1$ (53.6), the ranges widen to 3.2–5.1 points, and the $\lambda = 2$ run drifts from 51.2 to 46.2 within the window, so the coefficient that improves RIDE degrades ExOPD and makes its late-training behavior less stable. For RIDE, every $\lambda \geq 1.15$ shortens responses relative to OPRD while $\lambda \leq 0.75$ lengthens them; for ExOPD, the final Avg@16 is non-monotone in λ and every $\lambda > 1$ ends below the $\lambda = 1$ run, and the format score decreases with λ except for a partial recovery at $\lambda = 1.5$.

Table 2: Extrapolation-coefficient sweep on the R1-Distill-1.5B pair (one run per coefficient). Per-benchmark and Avg@16 columns are validation Avg@16 (%) at the final checkpoint; "Format" is the validation format score (%); "Length" is the mean response length (tokens) over the 50 steps preceding the final checkpoint. Bold marks the best final Avg@16 within each method.

Method	λ	AIME24	AIME25	AIMO	Avg@16	Format	Length
RIDE	0.5	42.4	28.2	74.0	48.2	85.3	7430
	0.75	49.6	32.5	79.6	53.9	96.0	6789
	1 (OPRD)	46.5	35.0	81.3	54.3	97.2	5726
	1.15	48.8	36.0	80.9	55.2	96.4	5663
	1.25	49.8	35.6	80.9	**55.4**	96.6	5569
	1.35	49.8	34.8	81.4	55.3	94.2	5570
	1.5	46.5	32.9	77.6	52.3	77.1	5514
	2	43.8	34.0	79.4	52.4	86.4	5325
ExOPD	0.5	44.2	30.6	75.8	50.2	92.1	7032
	0.75	47.1	33.8	78.0	**52.9**	84.7	6816
	1 (OPD)	47.1	33.5	78.1	**52.9**	80.9	6998
	1.25	44.2	29.6	75.8	49.9	76.0	7404
	1.5	42.9	32.3	76.2	50.5	83.7	6300
	2	37.2	28.7	72.6	46.2	64.4	8704

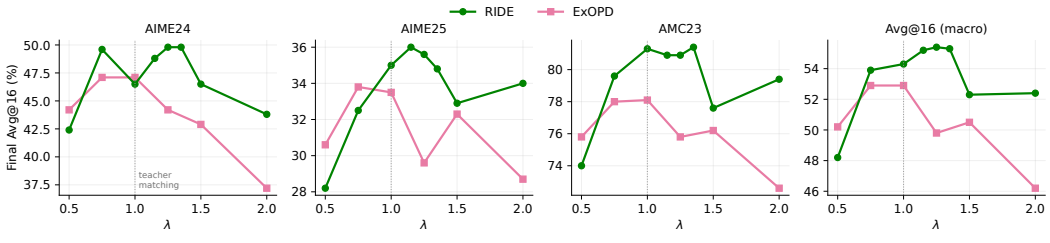


Figure 6: Final Avg@16 (%) per benchmark as a function of the extrapolation coefficient on the R1-Distill-1.5B pair (one run per coefficient). The dotted line marks $\lambda = 1$, which is OPRD for RIDE and sampled-token OPD for ExOPD.

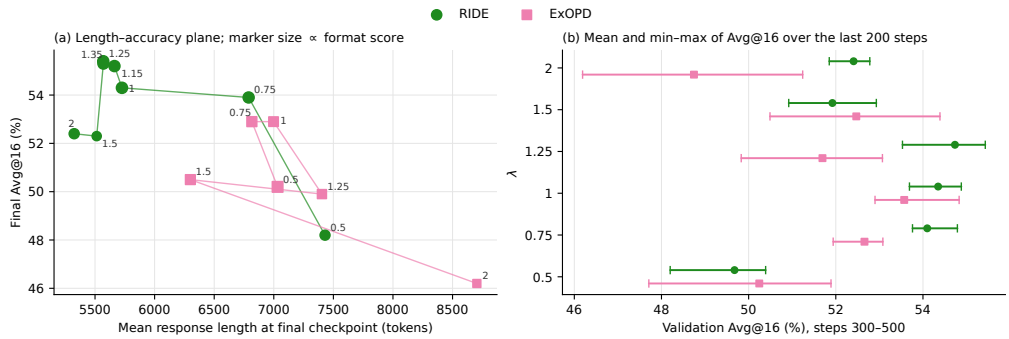


Figure 7: Extrapolation-coefficient sweep viewed per run. (a) Final Avg@16 against mean response length at the final checkpoint; markers are labeled by λ, connected in increasing λ, and sized by validation format score. (b) Mean (marker) and min–max range (bar) of validation Avg@16 over the last 200 training steps (steps 300–500) for the six coefficients tracked in Fig. 4.

C Direction Controls

Table 3 lists the direction-control results discussed in Section 5.5. Every control keeps the training protocol, the displacement magnitude $|\lambda - 1| \, \|\Delta_t^{(l)}\|$, and the layer and position sets of RIDE fixed and changes only the direction of the displacement.

Table 3: Direction controls at fixed displacement magnitude; Avg@16 (%) averaged over AIME24, AIME25, and AIMO on the R1-Distill-1.5B pair; $\lambda = 1.25$ unless stated.

Target construction	Avg@16
Teacher matching (OPRD, $\lambda = 1$)	54.50
Random direction, same norm	55.04
Reversed direction ($\lambda = 0.75$)	53.90
Mismatched origin ($h_T - h_{B'}$)	54.75
Trajectory-mismatched residual	55.12
RIDE (RL-induced residual)	56.38

D Training-Seed Summary

Table 4 summarizes the main-table macro Avg@16 means and across-seed standard deviations for OPRD and RIDE on the four base/RL-teacher pairs. Each run-level macro score averages the three benchmark Avg@16 scores with equal weight, as in Section 5.1. The three independent training seeds are 14, 42, and 2027.

Table 4: Mean $\pm$ SD of macro Avg@16 (%) over three independent training runs (seeds 14, 42, and 2027). Means are taken from Table 1; SDs are across training seeds. Teacher scores are fixed-checkpoint references.

Base/RL-teacher pair	Teacher	OPRD	RIDE
R1-Distill-1.5B	55.30	54.50 ± 1.04	56.38 ± 0.97
Qwen3-4B	65.59	62.01 ± 0.96	66.07 ± 0.88
Llama-3.2-3B	13.01	10.94 ± 0.90	13.33 ± 1.13
Phi-4-mini	17.96	17.33 ± 1.06	18.30 ± 1.05

E Additional Head Diagnostics

The measurements in Fig. 5 use 292,864 positions from the R1-Distill-1.5B pair: 143 problems, 16 rollouts each, and 128 positions per rollout from the last 2,000 response tokens at step 500. The residual's normalized head gain is 0.59 (prompt-bootstrap 95% CI $[0.585, 0.596]$), and its hidden-state energy outside the top-r head subspace exceeds the isotropic expectation at every r. The centered head is full rank, with singular values ranging from 71.70 to 1.78, so the residual is attenuated rather than removed. Across training steps, the 512 weakest directions carry 73.00% of the representation-loss gradient of Eq. (7), versus 48.60% of the output-KL gradient at the head input.

Scope of Eq. (8). Equation (7) supervises the pre-normalization residual stream, so Eq. (8) holds after the final normalization is applied to the target; the identity is stated at the head input so that it holds exactly. It also assumes the head shared by base and student: a teacher whose head drifted during RL contributes an additional term $(W_T - W_B)\tilde{h}_T$ to its native output-space target, quantified below. With W_B fixed, the head image of the RIDE target matches $\lambda z_T + (1-\lambda) z_B$ to relative error 6.50×10^{-7}. During RL, the teacher head drifts from W_B by 1.65% in relative Frobenius norm. The resulting term $(W_T - W_B)\tilde{h}_T$ has 50% of the norm of the native logit change, compared with 94% for the representation term, and the two have cosine similarity -0.23. Fixed-head and native-head targets differ by 18.70% in mean relative logit error, although the corresponding distributions have a mean KL divergence of 0.0058 and a top-1 agreement of 96.90%. The representation-space target contains no head-drift term. For the student update in Fig. 5(b), the residual-alignment cosine is 0.954 (prompt-bootstrap 95% CI $[0.949, 0.958]$), the head-visible cosine is 0.895 over 5.20×10^5 coordinates, and the mean projection along the residual is 1.69 (CI $[1.65, 1.73]$).

F Additional Learning-Signal Diagnostics

From each method's on-policy validation rollouts, we use 143 problems, 4 rollouts per problem, and 16 positions per rollout drawn uniformly from the last 2,000 response tokens, giving 9,152 fixed prefixes per method. For ExOPD, summing over the full vocabulary of 151,936 tokens at each prefix gives the exact conditional moments $\mathrm{Var}[\ell]$, $\mathrm{Var}[\rho]$, and $\mathrm{Cov}[\ell, \rho]$, and hence $\mathrm{Var}_{v\sim p}[A_\lambda]$ for any λ. Resampling the next token 8,192 times per prefix reproduces the exact values to within 0.38%. For RIDE, we measure $\|\nabla_h \mathcal{L}_\lambda\|_2$ at each fixed prefix and resample the next token 8,192 times per prefix (7.50×10^7 draws in total, with adjacent draws differing in 19.50% of cases).

Averaged over prefixes, the ExOPD conditional variance follows Eq. (9) with fitted coefficients

$$\mathrm{Var}[A_\lambda] \;=\; 0.0096 \;-\; 0.0421\,(\lambda - 1) \;+\; 0.153\,(\lambda - 1)^2.$$

The terms are $\mathrm{Var}[\ell] = 0.0096$, $-2\,\mathrm{Cov}[\ell, \rho] = -0.0421$, and $\mathrm{Var}[\rho] = 0.153$. The negative covariance term produces a minimum of 0.0067 at $\lambda \approx 1.14$; beyond this point the quadratic term dominates, giving $2.80\times$ the $\lambda = 1$ variance at $\lambda = 1.50$ and $12.60\times$ at $\lambda = 2$ (0.121 versus 0.0096; prompt-bootstrap 95% CIs $[0.101, 0.146]$ and $[0.0084, 0.0111]$, respectively). In Fig. 5c, shading shows these confidence intervals and the dashed curve isolates $(\lambda - 1)^2\,\mathrm{Var}[\rho]$, which accounts for essentially all growth beyond $\lambda \approx 1.30$.

For RIDE, the target $h^\star$ depends on the prefix and the frozen models but not on the token drawn next. The variance of its per-position gradient norm across resampled tokens is therefore exactly zero at every prefix and every λ; the red curve is placed at the plotting floor and does not represent a measured positive value. The gradient norm does vary across prefixes (between-prefix variance $\approx 4 \times 10^3$ at every λ), which reflects variation in the data rather than next-token sampling noise.

G Implementation and Diagnostics

Training and evaluation configuration. Table 5 lists the configuration shared by all methods; only the target construction differs between runs. Final answers are extracted with the standard boxed parser and graded by exact match. Table 6 summarizes the per-rollout cost of each method relative to OPRD: RIDE adds one frozen forward pass of π_B on the teacher worker and no additional communication, since $h^\star$ is formed in place and a single hidden-state tensor is transmitted, as for OPRD.

Table 5: Training and evaluation configuration shared by all methods and all four base/teacher pairs.

Rollout	
Prompt dataset	DAPO-Math-17K
Prompts per step / responses per prompt	8 / 2
Sampling temperature	1.00
Maximum response length	16,384 tokens
Optimization	
Steps	500
Optimizer / peak learning rate	AdamW / 1×10^{-5}
Schedule	3% linear warm-up, cosine decay
Precision / parallelism	bf16 / FSDP (verl)
Hardware	4× H200
RIDE target	
Supervised layers / positions	all L layers / last $k = 2{,}000$ positions
Extrapolation coefficient λ	1.25 (sweep in Section B)
Loss coefficient	scaled by λ^{-2} (Remark G.1)
Reference checkpoint	pre-RL base π_B (frozen)
Evaluation	
Metric	Avg@16, temperature 0.70
Benchmarks (problems)	AIME24 (30), AIME25 (30), AIMO (83; AMC 2022–2023)

Table 6: Per-rollout cost of each method. "Frozen passes" counts forward passes of frozen models on the student's rollout; "Reference" is whether the pre-RL checkpoint must be loaded; "To trainer" is the supervision tensor transmitted from the teacher worker per rollout (V: vocabulary size; L, d: supervised layers and hidden width).

Method	Frozen passes	Reference	To trainer	Target computed
OPD top-1 / top-16	1 (teacher)	no	$T \times 1$ / $T \times 16$ log-probs	on trainer
OPRD	1 (teacher)	no	$T \times L \times d$ hidden states	on teacher worker
ExOPD	2 (teacher, base)	yes	$T \times 1$ log-ratios	on trainer
RIDE	2 (teacher, base)	yes	$T \times L \times d$ hidden states	on teacher worker

Target construction. The frozen pre-RL checkpoint is loaded into the worker that already hosts the teacher. For each rollout, this worker runs one additional forward pass of π_B, forms $h^\star$ in place according to Eq. (6), and transmits a single hidden-state tensor to the trainer, so the transfer volume equals that of OPRD. The loss, layer alignment, and position mask are those of OPRD; positions outside the mask or beyond the response length are zero-padded in both hidden-state tensors, so $h^\star$ vanishes there and the masked loss is unaffected. When $\lambda = 1$ or no base checkpoint is provided, the base model is neither loaded nor evaluated, and training coincides with OPRD.

Remark G.1 (Loss scale). At initialization the student coincides with the base, $h_{\theta_0} = h_B$, so $h_{\theta_0} - h^\star = -\lambda\Delta$ and the initial per-position loss is $\lambda^2\|\Delta\|_2^2/d$, a factor λ^2 larger than for OPRD. Because Eq. (4) is unnormalized, changing λ therefore also changes the effective step size. We keep the unnormalized form so that $\lambda = 1$ recovers OPRD exactly, and multiply the loss coefficient by λ^{-2}, which equalizes the initial loss with that of OPRD for every λ; loss values at different λ are compared after this rescaling.

Diagnostics. During training we log the cosine similarity between h_B and h_T, the residual norm ratio $\|\Delta\|_2/\|h_T\|_2$ both overall and per layer, and the target norm ratio $\|h^\star\|_2/\|h_T\|_2$, which directly monitors the scale sensitivity described in Remark G.1.