STATE-AWARE CONTRASTIVE ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation trains a student on its own generations, while a single-teacher objective specifies only the distribution to imitate and no explicit extrapolation direction. We propose *Contrastive On-Policy Distillation* (COPD), which combines positive-teacher attraction and negative-teacher repulsion in a conditional KL objective on student-generated prefixes. We derive its closed-form virtual teacher and characterize the induced targets as an exponential-family path. The teacher pair supplies both an extrapolation direction and a coordinate system for targets and students. We localize the student along this path with a one-dimensional convex reverse-KL projection, whose residual measures off-axis mismatch. We extend COPD with *State-Aware Contrastive On-Policy Distillation* (SACOPD), which places each rollout's target a fixed coordinate increment ahead of the student's projection. With a Qwen3-4B student and 300 training rollouts, static COPD raises mathematics sample accuracy from matched OPD's 46.17% to 58.91%, compared with the positive teacher's 49.04%. SACOPD reaches 59.48% and improves both accuracy and coverage over COPD on all three code benchmarks. Together, these results connect effective teacher amplification with interpretable student localization and state-aware target selection.

1 INTRODUCTION

On-policy distillation (OPD) is a direct way to transfer capabilities between language models. The student samples responses from its own policy, and the teacher supplies dense token-level supervision on those same trajectories. This places training on the states the student will actually visit at inference time and reduces the mismatch introduced by teacher-forcing distillation (Gu et al., 2024; Agarwal et al., 2024). Recent work has made OPD increasingly effective through improved divergences, rollout allocation, and teacher–student compatibility (Lu & Lab, 2025; Ye et al., 2026b; Ziheng et al., 2026; Li et al., 2026).

Fixed-teacher OPD applies supervision on-policy while retaining a static matching target. A single teacher specifies a distribution to imitate. A teacher pair can additionally specify a policy change to amplify through its positive–negative log-probability difference. This motivates two connected questions: how should that contrast define supervision beyond the positive teacher, and where should the target lie relative to the evolving student?

We propose Contrastive On-Policy Distillation (COPD).[1] COPD combines positive-teacher attraction and negative-teacher repulsion in a two-sided conditional KL objective on student-generated prefixes. Its local virtual teacher defines an exponential-family path of targets, providing a common geometry for contrastive supervision and student state.

We identify two central roles for a teacher pair: a direction and a ruler. Its log-probability difference defines a path through the teacher anchors to extrapolated targets. A convex reverse-KL projection locates the student on this path, yielding a signed coordinate that remains interpretable beyond the positive teacher and a residual measuring off-axis mismatch. State-Aware Contrastive On-Policy Distillation (SACOPD) repeats this localization on each rollout and analytically places its target a fixed coordinate increment ahead.

[1] Our formulation was developed independently of the concurrent work of Ruan et al. (2026), which uses the same name and acronym; we discuss their relationship in Section 2.

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.

On Qwen3-4B mathematical reasoning under a competition benchmark with 32 samples per problem and a 32K-token generation budget, COPD raises mean@32 from matched OPD's 46.17% to 58.91%, compared with the positive teacher's 49.04%; SACOPD further improves mean@32 to 59.48%. On code generation, COPD improves mean@4 from OPD's 66.79% to 67.98%, and SACOPD extends it to 69.77%. SACOPD raises both accuracy and coverage over COPD on HumanEval+, MBPP+, and LiveCodeBench, with mean@4 gains of 2.59, 2.58, and 0.19 percentage points, respectively. Coordinate audits additionally resolve student movement along the teacher axis during an accuracy plateau, connecting the target geometry to an inspectable optimization state.

Our contributions are:

- We propose COPD, formalizing positive-teacher attraction and negative-teacher repulsion in a two-sided conditional KL objective. Its closed-form virtual teacher gives contrast strength an explicit interpretation as position on a teacher-defined target path.
- We derive a convex reverse-KL projection that maps the student distribution onto this teacher-defined axis, yielding a signed optimization coordinate and a residual that distinguishes position from off-axis mismatch.
- We extend COPD with SACOPD, a state-aware controller that reads the projected student coordinate and sets contrast strength to place the virtual teacher a fixed increment ahead, closing the target-selection loop without additional task rewards or a learned controller.
- On Qwen3-4B, we evaluate static amplification, state-aware target selection, and coordinate dynamics. SACOPD improves accuracy and coverage over COPD on all three code benchmarks, while fixed-replay audits track student position and projection quality.

2 Related Work

2.1 On-policy distillation

On-policy distillation supervises students on their own generated prefixes, reducing the training–inference distribution mismatch of teacher-forced distillation. The reverse-KL variant concentrates a capacity-limited student on the teacher's high-probability modes (Gu et al., 2024); GKD frames the process as imitation learning with an interactive expert and separates the choices of data mixture and divergence (Agarwal et al., 2024). Subsequent work refines supervision through entropy-aware divergence weighting (Jin et al., 2026), log-ratio rewards, and early-stopped rollouts (Ko et al., 2026; Ziheng et al., 2026), or varies teacher conditioning through privileged-context and cross-modal distillation (Ye et al., 2026b;a; Cao et al., 2026). A complementary direction constructs supervision targets beyond the teacher distribution.

2.2 Extrapolative and contrastive OPD

Extrapolative supervision uses a reference or teacher pair to define a direction beyond the teacher. Generalized OPD (Yang et al., 2026) derives this construction from KL-regularized reward maximization, with ExOPD scaling the teacher–reference reward above one; COPD starts from positive-teacher attraction and negative-teacher repulsion in a two-sided conditional KL.

The use of teacher differences also connects COPD to weak-to-strong transfer and prompt-conditioned distillation. W2S-OPD (Yu et al., 2026) transfers weak teachers' capability differences to a stronger student by adding their scaled logit contrast to a frozen student-base anchor. For concise reasoning, Ruan et al. (2026) introduce a prompt-conditioned method also named COPD, contrasting a frozen teacher's scores under light/heavy reasoning prompts, including a self-distillation variant. Their contrast signal formally matches the $\lambda{=}1$ boundary of our two-sided KL and is optimized with clipped policy gradients. Our work was developed independently and concurrently with these two studies.

COPD formulates attraction and repulsion through a two-sided KL objective for capability amplification with general teacher pairs. We characterize its closed-form next-token virtual teacher and extrapolation geometry in the $0{<}\lambda{<}1$ regime, giving contrast strength an explicit interpretation as target position on a teacher-defined path. This same path further supports student localization and off-axis mismatch analysis through reverse-KL projection. Building on COPD, SACOPD places

each rollout's target a fixed increment ahead of the projected student position, extending the formulation to state-relative target selection.

3 Contrastive Distillation and State-Aware Targets

3.1 Contrastive On-Policy Distillation

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the prompt and response spaces, respectively, and let $\mathcal{D}$ be the prompt distribution. We train an autoregressive student π_θ using a fixed positive teacher π^+ and a fixed negative teacher π^-. For a response $y = (y_1, \ldots, y_{|y|})$, every policy factorizes as $\pi(y \mid x) = \prod_t \pi(y_t \mid y_{<t}, x)$. At update k, the current student snapshot $\bar{\theta}_k$ generates responses under the training rollout protocol for $x \sim \mathcal{D}$. Let ν_k denote the empirical distribution of prefixes $h = (y_{<t}, x)$, with weights matching the training loss aggregation. During this update, ν_k is fixed and the conditional student $\pi_\theta(\cdot \mid h)$ remains differentiable.

For reference, reverse-KL OPD with a single teacher is the fixed-prefix conditional objective

$$\mathcal{L}^{(k)}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{h\sim\nu_k}\left[\mathrm{KL}\big(\pi_\theta(\cdot \mid h) \,\big\|\, \pi^+(\cdot \mid h)\big)\right]. \tag{1}$$

Over unrestricted conditional distributions, its pointwise minimizer is the positive teacher. The objective therefore specifies a matching target, but no explicit direction for extrapolation beyond that target. To introduce that direction, we propose Contrastive On-Policy Distillation (COPD), which combines attraction to a positive teacher with repulsion from a negative teacher:

$$\boxed{\mathcal{L}^{(k)}_{\mathrm{COPD}}(\theta) = \mathbb{E}_{h\sim\nu_k}\left[\mathrm{KL}\big(\pi_\theta \,\big\|\, \pi^+\big) - \lambda\,\mathrm{KL}\big(\pi_\theta \,\big\|\, \pi^-\big)\right], \qquad 0 < \lambda < 1.} \tag{2}$$

Both KLs compare next-token distributions at the same fixed prefix h. The restriction $0 < \lambda < 1$ preserves a positive coefficient on the conditional negative-entropy term. Training differentiates this objective with ν_k fixed, then refreshes the rollouts for the next update. Appendix A.9 gives the separate sequence-level result.

3.2 Local Optimum: Amplification via a Virtual Teacher

The conditional objective has an exact next-token target.

Proposition 1 (Closed-form local COPD optimum)**.** *For any fixed prefix h and $0 < \lambda < 1$, assume a finite vocabulary and strictly positive teacher probabilities. Over unrestricted next-token distributions, the unique minimizer of the integrand in* (2) *is*

$$\pi^\star_{\mathrm{local}}(v \mid h) = \frac{1}{Z_h}\pi^+(v \mid h)^\alpha \pi^-(v \mid h)^{1-\alpha}, \qquad \alpha = \frac{1}{1-\lambda} > 1, \tag{3}$$

where Z_h normalizes the next-token distribution.

With ν_k fixed, the objective has the exact per-update identity

$$\mathcal{L}^{(k)}_{\mathrm{COPD}}(\theta) = (1-\lambda)\mathbb{E}_{h\sim\nu_k}\,\mathrm{KL}(\pi_\theta(\cdot \mid h) \,\|\, \pi^\star_{\mathrm{local}}(\cdot \mid h)) + C_k, \tag{4}$$

where $C_k = -(1-\lambda)\mathbb{E}_{h\sim\nu_k}\log Z_h$ is independent of the updated θ (proof in Appendix A.1). Defining $\Delta(v \mid h) = \log\pi^+(v \mid h) - \log\pi^-(v \mid h)$, we can rewrite the target as

$$\log\pi^\star_{\mathrm{local}}(v \mid h) = \log\pi^+(v \mid h) + (\alpha-1)\Delta(v \mid h) - \log Z_h. \tag{5}$$

The target extrapolates the positive–negative log-probability shift with coefficient $\alpha - 1 = \lambda/(1-\lambda)$. When $\lambda \to 0$, it reduces to π^+ and COPD recovers one-sided OPD. Sections 3.3–3.5 develop target geometry and student localization for these conditional distributions.

3.3 Contrast Strength as Target Position

For a response-prefix context $h = (y_{<t}, x)$ and vocabulary token v, define the token-level contrast direction

$$\Delta(v \mid h) = \log\pi^+(v \mid h) - \log\pi^-(v \mid h). \tag{6}$$

The teacher pair induces the exponential-family path

$$\pi_\rho(v \mid h) = \pi^+(v \mid h) \exp(\rho\, \Delta(v \mid h) - A_h(\rho)), \qquad A_h(\rho) = \log \sum_u \pi^+(u \mid h)\, e^{\rho \Delta(u|h)}. \quad (7)$$

Two points have an intrinsic interpretation:

$$\pi_{-1}(\cdot \mid h) = \pi^-(\cdot \mid h), \qquad \pi_0(\cdot \mid h) = \pi^+(\cdot \mid h). \quad (8)$$

The coordinate increment from the negative anchor to the positive teacher is one axis unit. Positive values of ρ extrapolate the same teacher log-ratio beyond the positive anchor. The local virtual-teacher optimum $\pi^\star_{\text{local}}(\cdot \mid h) = \pi_{\alpha-1}(\cdot \mid h)$ with $\alpha - 1 = \lambda/(1-\lambda)$ thus corresponds to coordinate $\rho = \alpha - 1 > 0$ on this path.

To parameterize contrast strength while keeping the endpoint teachers fixed, define a controlled negative

$$\widetilde{\pi}^-_\beta(\cdot \mid h) = \pi_{-\beta}(\cdot \mid h) \propto \pi^+(\cdot \mid h)\, e^{-\beta \Delta(\cdot|h)}, \qquad \beta \geq 0. \quad (9)$$

Here $\beta = 1$ recovers the natural base negative π^-, while intermediate and extrapolated values move continuously along the same axis. Applying the COPD objective with fixed λ to $(\pi^+, \widetilde{\pi}^-_\beta)$ yields a virtual optimum on the forward path at

$$\pi_{\rho_{\lambda,\beta}}(\cdot \mid h), \qquad \rho_{\lambda,\beta} = \frac{\lambda\beta}{1-\lambda}. \quad (10)$$

Thus β and λ are implementation parameters, whereas $\rho_{\lambda,\beta}$ is the location of the target distribution on the shared teacher axis.

3.4 Localizing the Student on the Teacher Axis

The teacher path assigns coordinates to the anchors and virtual targets, but the student need not belong to this one-dimensional family. We therefore define its position through a reverse-KL projection evaluated on the prefixes of each student-generated rollout.

For a rollout sample s, let $\mathcal{H}_s = \{h_t\}_{t=1}^{T_s}$ be the prefix contexts used for localization. We project the current student onto the teacher path:

$$\widehat{\rho}_s = \arg\min_{\rho \in \mathbb{R}} \Phi_s(\rho), \qquad \Phi_s(\rho) = \frac{1}{T_s} \sum_t \mathrm{KL}(\pi_\theta(\cdot \mid h_t) \,\|\, \pi_\rho(\cdot \mid h_t)). \quad (11)$$

The projection uses conditional distributions on a common token space and accommodates students with different parameterizations and initializations from the teachers.

Let $F_s(\rho) = T_s^{-1} \sum_t A_{h_t}(\rho)$ and $m_s = T_s^{-1} \sum_t \mathbb{E}_{\pi_\theta(\cdot|h_t)}[\Delta(\cdot \mid h_t)]$.

Proposition 2 (Existence and uniqueness of the projection)**.** *Assume a finite vocabulary, common strictly positive teacher support, and at least one prefix with nonconstant $\Delta(\cdot \mid h)$. If the student's average contrast moment m_s lies in the interior of the feasible range (m_-, m_+), the projection objective is strictly convex and admits a unique finite minimizer satisfying*

$$F_s'(\widehat{\rho}_s) = m_s. \quad (12)$$

Proof sketch. $F_s''(\rho) > 0$ at every finite ρ under the support assumption, so F_s' is strictly increasing and continuous; the intermediate value theorem gives the unique root. Full proof and boundary-case discussion are in Appendix A.3.

The resulting signed coordinate gives the student's position relative to teacher anchors. In particular, a student that has crossed the positive-teacher coordinate retains a positive projection value, while a single reverse-KL to π^+ cannot separate this case from one where the student is simply close to π^+.

We additionally record the projection residual $R_s = \Phi_s(\widehat{\rho}_s)$ and local curvature $I_s = \Phi_s''(\widehat{\rho}_s)$. The residual measures distributional mismatch not captured by the teacher axis; curvature characterizes the local geometry of the projection objective. Their precise relationship to the student–target distance is derived in Appendix A.4.

3.5 State-Aware Target Selection

SACOPD places each rollout's target a fixed coordinate increment ahead of the student:

$$\rho_s^{\text{target}} = \widehat{\rho}_s + \delta \tag{13}$$

By (10), the controlled-negative strength is

$$\beta_s = \frac{1-\lambda}{\lambda}(\widehat{\rho}_s + \delta). \tag{14}$$

On each update, the student is localized on its rollout prefixes to determine per-sample targets and strengths, which are detached during optimization. The conditional objective (2) then uses each rollout's controlled negative at its sampled prefixes. Repeating this procedure adapts supervision throughout training while keeping the endpoint teachers fixed.

If the student initializes at the negative anchor, $\pi_{\theta_0}(\cdot \mid h) = \pi^-(\cdot \mid h)$, then $\widehat{\rho}_{s,0} = -1$ and $R_{s,0} = 0$ on an identifiable axis. The nonnegative-strength domain $\beta_s \geq 0$ requires $\widehat{\rho}_s + \delta \geq 0$.

What the lookahead controls. A fixed δ advances the target one step along the teacher-axis coordinate; curvature and off-axis residual jointly determine the resulting student–target gap. The loss operates on full-vocabulary next-token distributions, with the projected coordinate determining target selection.

Parameterization equivalence. When prefixes, teachers, and rollout-specific control quantities are held fixed, the two-term conditional KL objective equals $(1-\lambda)$ times a single KL against a teacher-axis target, plus a term independent of the updated parameters (Appendix A.6). The equivalent target lies at the selected coordinate on the teacher path.

Teacher-pair design. A useful teacher pair shares general behavior but differs in the target capability, concentrating the contrast on that capability. The formulation applies to any fixed pair; gradient specificity and pair instantiations are discussed in Appendices A.7 and A.8.

4 Experimental Setup

Models and training. We initialize the student from Qwen3-4B (denoted Qwen3-4B Base, or Base, before task-specific RL) and use a frozen copy as the negative teacher. Our positive teachers are the publicly released Qwen3-4B domain experts provided by the authors of G-OPD (Yang et al., 2026): the Math-RL step-500 checkpoint for mathematics and the Code-RL step-300 checkpoint for code generation. Both were obtained through task-specific GRPO training from the same Qwen3-4B Base initialization. COPD and SACOPD use $\lambda = .80$ for mathematics and .20 for code; COPD uses the natural Base negative ($\beta = 1$), while SACOPD uses $\delta = 2$ and 1, respectively. The trained OPD, COPD, and SACOPD run for 50 updates with global batch size six and use a 256-token response window for training and projection. Training configurations and checkpoints are reported in Appendix B.

Training datasets. For mathematics, we filter DeepMath-103K (He et al., 2026) to examples with difficulty level at least six, forming a 57K-sample math RL pool. For code, we use the 25K-sample Eurus-RL-Code dataset (Cui et al., 2025).

External reproductions. We reproduce ExOPD, the reward-extrapolation regime of Generalized OPD (Yang et al., 2026), and IW-OPD (Xie et al., 2026) using their released objectives on the same student, teacher, and training-data axes. The main table uses a matched 50-update, global-batch-six budget for these objectives; a released-scale batch-126 reproduction is reported separately in Appendix B.4.2 to retain the comparison with the released training scale. We also adapt W2S-OPD (Yu et al., 2026) to the same 4B axes and 300-rollout budget (Appendix B.4.1).

Evaluation. The mathematical evaluation combines AIME 2024, AIME 2025, HMMT February 2025, and HMMT November 2025, with 30 problems from each source. The primary comparison draws 32 responses per problem with a 32,768-token response cap. Math-Verify checks the final complete `\boxed{}` expression in the full response, and missing or malformed boxes are incorrect without an extraction fallback. Code is evaluated at four samples per problem on HumanEval+ and

MBPP+, the EvalPlus extensions of HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), and the February–May 2025 subset of LiveCodeBench v6 (Jain et al., 2025) used by G-OPD (Yang et al., 2026) (131 problems). HumanEval+/MBPP+ use the official EvalPlus (Liu et al., 2023) base-plus criterion and a 16,384-token response cap; LiveCodeBench uses the official execution scorer and a 32,768-token cap. All primary evaluations use non-thinking generation with temperature and top-p equal to one. Appendix B.3 specifies the custom extraction and LiveCodeBench prompt policies.

Metrics and supporting studies. For K samples per problem, `mean@`K is the fraction of correct responses and `pass@`K is the fraction of problems solved by at least one response. Parameter screens, teacher-pair comparisons, coordinate audits, and exposure curves use the sample counts stated with their tables; their complete grids and protocol differences are collected in Appendix B.

5 Results

5.1 Overall Results

Table 1: Mathematics (top) and code (bottom) results (%). m@K and p@K denote mean@K and pass@K; Overall is the within-domain macro-average. The trained comparison rows use 50 updates and 300 rollouts in this matched-budget panel. W2S-OPD configurations are in Appendix B.4.1; the released-scale external reproductions are reported separately in Appendix B.4.2. Bold marks column maxima; shading marks our methods.

Method	AIME24		AIME25		HMMT-F		HMMT-N		Overall	
	m@32	p@32	m@32	p@32	m@32	p@32	m@32	p@32	m@32	p@32
Student init	21.88	53.33	19.90	43.33	10.31	30.00	8.85	40.00	15.23	41.67
Positive teacher	62.40	86.67	58.54	76.67	34.48	63.33	40.73	70.00	49.04	74.17
Matched OPD	60.10	**90.00**	54.06	80.00	32.92	60.00	37.60	63.33	46.17	73.33
W2S-OPD	61.04	83.33	56.67	**83.33**	33.65	66.67	39.58	66.67	47.73	75.00
W2S-OPD ($\alpha = 5$)	**74.58**	**90.00**	67.08	**83.33**	42.50	60.00	50.31	76.67	58.62	77.50
IW-OPD	62.19	83.33	58.02	86.67	40.94	**70.00**	36.56	66.67	49.43	76.67
ExOPD	68.44	90.00	59.69	83.33	38.65	66.67	44.58	66.67	52.84	76.67
COPD	73.65	86.67	68.44	**83.33**	42.50	**70.00**	**51.04**	76.67	58.91	79.17
SACOPD	73.65	**90.00**	**69.48**	**83.33**	**43.85**	**70.00**	50.94	**80.00**	**59.48**	**80.83**

Method	HE+		MBPP+		LCB		Overall	
	m@4	p@4	m@4	p@4	m@4	p@4	m@4	p@4
Student init	79.73	93.29	65.94	74.87	23.85	31.30	56.51	66.49
Positive teacher	**91.01**	**96.34**	71.10	82.01	40.08	46.56	67.39	74.97
Matched OPD	86.59	93.90	73.15	82.80	40.65	51.15	66.79	75.95
W2S-OPD	84.45	94.51	69.18	79.63	33.78	41.98	62.47	72.04
W2S-OPD ($\alpha = 5$)	87.35	92.68	**76.72**	**84.39**	42.18	50.38	68.75	75.82
IW-OPD	87.65	95.73	70.90	82.01	35.88	49.62	64.81	75.79
ExOPD	86.89	95.73	71.16	83.60	39.31	51.15	65.79	76.82
COPD	87.20	95.12	73.61	81.48	43.13	54.20	67.98	76.93
SACOPD	89.79	**96.34**	76.19	83.60	**43.32**	**54.96**	**69.77**	**78.30**

Static extrapolation supplies the main mathematical gain; state-aware targets extend it. COPD raises mean@32 from matched OPD's 46.17% to 58.91% (Table 1). Strength-matched W2S-OPD reaches 58.62%, supporting the shared fixed-target mechanism. SACOPD reaches 59.48% mean@32 and 80.83% pass@32. On code, SACOPD raises macro-average mean@4 from COPD's 67.98% to 69.77%, improving both accuracy and coverage on all three benchmarks. These comparisons establish the effectiveness of static amplification and motivate examining student-relative

target placement through selection controls and coordinate audits below. Under the matched 300-rollout budget, IW-OPD and ExOPD reach 49.43%/76.67% and 52.84%/76.67% on mathematics, respectively; the corresponding released-scale external results are retained in Appendix B.4.2. Both COPD and SACOPD retain their overall accuracy and coverage gains over OPD in two additional training seeds (Appendix B.6).

W2S-OPD (Yu et al., 2026) retains the paper's coefficients ($\alpha = 1$ for math, .75 for code). With our base anchor equal to the negative teacher, $\alpha = 1$ reduces to top-32 OPD without extrapolation. We additionally use $\alpha = 5$ in both domains, matching math COPD's virtual-target strength ($\lambda = .8$, $\beta = 1$); Appendix B.4.1 details the target correspondence and configurations.

5.2 Parameter Sensitivity

Lookahead δ. Table 2(a) varies lookahead at fixed $\lambda = .80$ under $K = 8$ evaluation. The controller columns summarize the final on-policy batch, and the audit column uses a common fixed replay (Appendix B.2). Larger lookaheads advance the target and increase contrast strength, while the replay-based student coordinate changes little for $\delta \geq 2$. Across all five lookaheads, mean@8 ranges from 58.44% to 59.79% (a 1.35-point span), while pass@8 increases from 71.67% to 75.83%. These results show a broad effective lookahead range in this setting.

Table 2: Sensitivity to lookahead δ and parameterization λ on mathematics. Bold marks the best performance within each panel; shading marks the main setting. The shaded $\lambda = .80$ row in (b) reuses the checkpoint and evaluation from the shaded $\delta = 2$ row in (a).

(a) Vary lookahead δ Fixed $\lambda = 0.80$

	Performance (%)		On-policy controller			Fixed replay
δ	mean@8	pass@8	Student $\widehat{\rho}$	Target ρ^{target}	Strength β	Audit $\widehat{\rho}$
1.0	**59.79**	71.67	1.215	2.215	0.554	1.284
1.5	59.58	71.67	1.324	2.824	0.706	1.347
2.0	59.17	73.33	1.391	3.391	0.848	1.405
2.5	58.65	73.33	1.390	3.890	0.972	1.405
3.0	58.44	**75.83**	1.351	4.351	1.088	1.408

(b) Vary parameterization λ Fixed $\delta = 2$

	Performance (%)		On-policy controller		
λ	mean@8	pass@8	Student $\widehat{\rho}$	Target ρ^{target}	Strength β
0.20	59.58	68.33	1.41	3.41	13.63
0.50	**59.79**	**74.17**	1.26	3.26	3.26
0.80	59.17	73.33	1.39	3.39	0.85
0.90	58.44	72.50	1.34	3.34	0.37

Parameterization λ. At fixed $\delta = 2$, tested settings $\lambda \in \{.20, .50, .80, .90\}$ span only 1.35 percentage points in mean@8, from 58.44% to 59.79% (Table 2(b)). Their final on-policy student coordinates range from 1.26 to 1.41, while resolved strengths β range from 0.37 to 13.63. These observations support practical role of coordinate-based target selection: different parameterizations can express similar target positions and achieve comparable single-sample accuracy. Pass@8 ranges from 68.33% to 74.17%, showing that answer coverage remains more variable across these runs.

5.3 Analysis

5.3.1 Influence of Teacher Pair

COPD amplifies both pre-/post-RL and cross-specialist teacher contrasts. Table 3 compares three teacher pairs with a fixed 4B student and COPD ($\lambda = .80$, $\beta = 1$), each against its matched OPD at $K = 8$ (Appendix B.5). Both Math-RL/Base and Math-RL/Code-RL yield substantial

Table 3: OPD and COPD across teacher pairs. Entries are mean@8 / pass@8 (%).

Teacher pair	OPD	COPD
Pre-/post-RL	45.73 / 65.83	59.58 / 74.17
Cross-specialist	45.73 / 65.83	59.69 / 76.67
Dense-size	15.73 / 30.83	4.90 / 14.17

Table 4: Performance under fixed, replay, and state-aware target selection (%).

Method	mean@32	pass@32
COPD	58.91	79.17
Global replay	58.88	80.00
SACOPD	59.48	80.83

gains, supporting teacher contrasts defined by either RL improvement or specialization. Choosing the negative teacher therefore determines which changes in the positive teacher are amplified; shared general behavior with an interpretable capability difference is a useful starting point.

Appendix B.5.1 reports the three teacher pairs with a 1.7B student, including a lower-strength dense-size configuration.

For the dense-size pair (Qwen3-14B/4B Base), a post-hoc change from $\lambda = .80$ to $\lambda = .20$ improves accuracy and reduces Cyrillic-character outputs from $391/960$ to $1/960$. This sensitivity is consistent with amplification of incidental teacher differences. SACOPD adapts target position along a given axis, but does not assess whether that axis isolates the intended capability.

5.3.2 From Fixed to State-Aware Target Selection

State-aware selection controls the target's lead over the student. Fixed COPD sets an absolute target coordinate, so its lead over the student shrinks as student advances and differs across rollouts. Global replay assigns a shared target at each training step using a donor run's mean schedule. SACOPD places each rollout's target a fixed coordinate increment ahead of current student's projection.

SACOPD achieves the highest observed accuracy and coverage among the three selection rules in Table 4, supporting target placement relative to student state. This formulation separates two design choices: the teacher pair defines the extrapolation direction, while student localization determines where to place supervision along it.

5.3.3 Contrast Strength

The natural base negative lies within a broad effective region of fixed strengths. At $\lambda = .80$, the 32K/$K = 8$ sweep in Figure 1 shows strong performance near $\beta = 1$, with the highest observed mean@8 at $\beta = 1.10$ (60.00%). SACOPD selects targets from this same family using student state, avoiding a grid search for one fixed β throughout training. Appendix C provides full results.

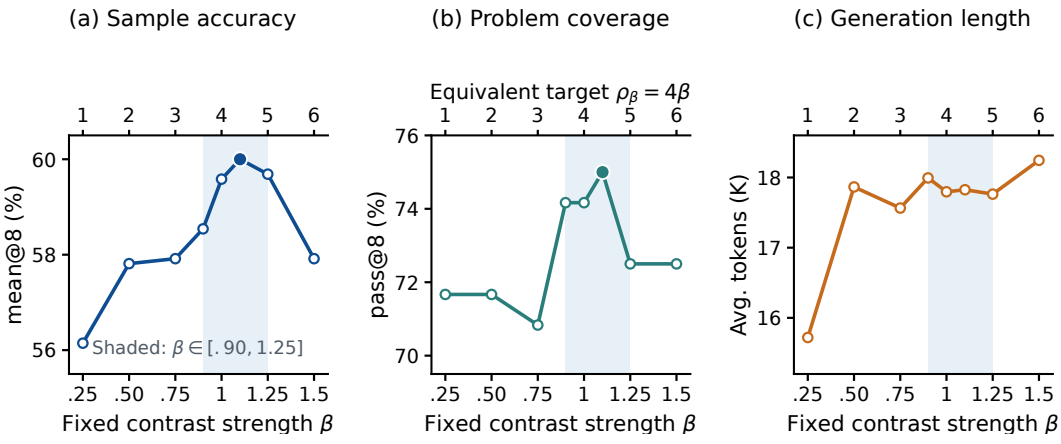


Figure 1: Mathematics performance versus fixed contrast strength β at 32K/$K = 8$. The upper axis shows target coordinate $\rho_{\text{target}} = 4\beta$.

5.3.4 Problem-Solving Behavior

Contrastive distillation supports sustained problem solving. SACOPD tests and verifies a lower-degree polynomial and develops an efficient square-root-domain sieve in the case studies (Appendix E), illustrating sustained analysis leading to verified solutions and executable algorithms.

5.4 Teacher-Axis Coordinate as an Optimization State

The teacher-axis coordinate resolves student progress relative to the teacher anchors throughout optimization. We audit a static COPD trajectory on the fixed Math-RL/Base axis using a common 128-response development replay and the first 256 response prediction prefixes (Appendix D). Holding these prefixes fixed isolates policy changes on common observed states. Figure 2 tracks coordinate evolution and projection quality through step 150.

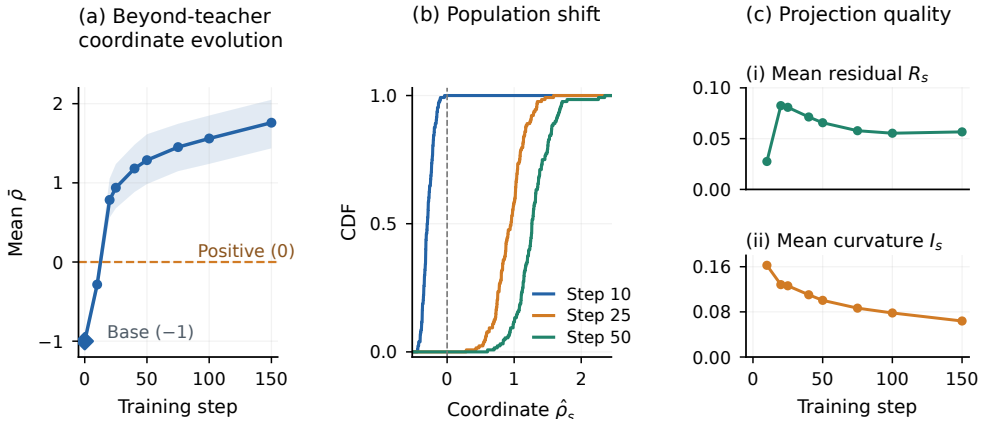


Figure 2: Teacher-axis coordinate diagnostics. **(a)** Mean coordinate and P10–P90 band; zero is the positive teacher. **(b)** Coordinate CDFs. **(c)** Mean residual and curvature across checkpoints.

The projection recovers both teacher anchors with near-zero residuals (Appendix D.2). All 128 replay coordinates cross from negative at step 10 to positive by step 20 and move forward in every subsequent observed interval through step 150 (Figure 2a–b; Appendix D.3). From step 20 to 150, the coordinate continues to resolve policy motion along the teacher axis: its mean advances from 0.7842 to 1.7597 while mean@4 stays within 57.71–60.21% (Appendix D.4).

Mean residual rises around the crossing, then declines; decreasing curvature indicates a flatter local projection objective (Figure 2c). These complement position with off-axis mismatch and local geometry (Appendix A.4). As student approaches fixed target coordinate of 4, the target's lead shrinks, motivating SACOPD's prescribed coordinate increment ahead of each rollout's student projection.

6 Conclusion

COPD formalizes positive-teacher attraction and negative-teacher repulsion in a two-sided conditional KL objective. Its virtual-teacher path gives targets and students a shared coordinate system. Reverse-KL projection locates the student on this axis and separates its position from off-axis mismatch; SACOPD uses that position to place supervision relative to the evolving student. Experiments with Qwen3-4B support static teacher amplification, while coordinate audits resolve policy movement during an accuracy plateau. SACOPD further improves accuracy and coverage over COPD on all three code benchmarks. This connects contrastive supervision to interpretable student localization and state-aware relative target placement.

AI USE STATEMENT

We used generative AI tools to assist with research execution, literature discovery, and manuscript preparation. Based on research ideas proposed by the authors, these tools helped develop concrete, feasible experimental designs and assisted in carrying out the experiments. They also supported the search for related work and the identification of relevant baseline methods for comparison, and helped draft and polish the manuscript. The authors manually reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, methodology, experimental results, claims, and associated artifacts.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Di Cao, Dongjie Fu, Hai Yu, Siqi Zheng, Xu Tan, and Tao Jin. X-opd: Cross-modal on-policy distillation for capability alignment in speech llms. *arXiv preprint arXiv:2603.24596*, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. 2026:138306–138322, 2026.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. 2025:58791–58831, 2025.

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. *arXiv preprint arXiv:2603.07079*, 2026.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

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

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. *Advances in neural information processing systems*, 36:21558–21572, 2023.

Kevin Lu and Thinking Machines Lab. On-policy distillation, 2025. https://thinkingmachines.ai/blog/on-policy-distillation.

Jiacheng Ruan, Jun Tang, Wenzhen Yuan, Ting Liu, Shuai Bai, Dayiheng Liu, Zhibo Yang, and Yuzhuo Fu. Contrastive on-policy distillation. *arXiv preprint arXiv:2607.19046*, 2026.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation. *arXiv preprint arXiv:2606.22600*, 2026.

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*, 2026.

Tianzhu Ye, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Online experiential learning for language models. *arXiv preprint arXiv:2603.16856*, 2026a.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026b.

Fangxu Yu et al. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026.

A Theoretical Derivations and Additional Analysis

A.1 Virtual-Teacher Identity: Proof of Proposition 1

At a fixed prefix h, define $g_h(v) = \log \pi^+(v \mid h) - \lambda \log \pi^-(v \mid h)$. For any next-token distribution p, the conditional loss in (2) is

$$\begin{aligned}\ell(p; h) &= \mathbb{E}_{v\sim p}[(1-\lambda)\log p(v) - g_h(v)] \\ &= (1-\lambda)\,\mathrm{KL}(p \,\|\, \pi^\star_{\mathrm{local}}(\cdot \mid h)) - (1-\lambda)\log Z_h, \end{aligned} \tag{15}$$

where $\pi^\star_{\mathrm{local}}(v \mid h) = Z_h^{-1}\exp(g_h(v)/(1-\lambda))$ and $Z_h = \sum_v \exp(g_h(v)/(1-\lambda))$. The finite vocabulary and strictly positive teacher probabilities ensure $0 < Z_h < \infty$. Since $\mathrm{KL} \geq 0$, with equality only at the target, this is the unique unrestricted conditional minimizer. Substituting $g_h/(1-\lambda) = \alpha \log \pi^+ + (1-\alpha)\log \pi^-$ proves (3).

Averaging over the fixed empirical prefix distribution ν_k gives (4). Both ν_k and Z_h are independent of the parameters being updated, so the identity holds for values and gradients, with the positive scale $(1-\lambda)$. A parameterized student attains the pointwise optimum only if it can realize the target on every prefix with positive ν_k weight; otherwise, the identity still characterizes its conditional KL fitting objective.

A.2 Coordinate Mapping in the Teacher Axis

Here we verify that the local virtual-teacher optimum $\pi^\star_{\mathrm{local}}$ defined in (3) corresponds to coordinate $\rho = \alpha - 1$ on the path (7).

From (3),

$$\begin{aligned}\log \pi^\star_{\mathrm{local}}(v \mid h) &= \alpha \log \pi^+(v \mid h) + (1-\alpha)\log \pi^-(v \mid h) + \mathrm{const} \\ &= \log \pi^+(v \mid h) + (\alpha - 1)\big[\log \pi^+(v \mid h) - \log \pi^-(v \mid h)\big] + \mathrm{const} \\ &= \log \pi^+(v \mid h) + (\alpha - 1)\Delta(v \mid h) + \mathrm{const}. \end{aligned} \tag{16}$$

Comparing with (7) shows that $\pi^\star_{\mathrm{local}}(\cdot \mid h) = \pi_{\alpha-1}(\cdot \mid h)$, confirming the coordinate assignment $\rho = \alpha - 1 = \lambda/(1-\lambda)$.

Applying the SACOPD objective with teacher pair $(\pi^+, \widetilde{\pi}^-_\beta)$ and fixed λ, the controlled negative $\widetilde{\pi}^-_\beta(\cdot \mid h) = \pi_{-\beta}(\cdot \mid h)$ enters the contrast direction as

$$\widetilde{\Delta}_\beta(v \mid h) = \log \pi^+(v \mid h) - \log \pi_{-\beta}(v \mid h) = \beta\Delta(v \mid h) + A_h(-\beta). \tag{17}$$

The log-partition term $A_h(-\beta)$ is independent of v and is absorbed into the normalization of the local virtual optimum. Substituting into the expression above gives

$$\log \pi^{\star}_{\text{local}}(v \mid h) = \log \pi^{+}(v \mid h) + \frac{\lambda\beta}{1-\lambda}\Delta(v \mid h) + \text{const.} \tag{18}$$

The resulting virtual optimum therefore sits at $\rho_{\lambda,\beta} = \lambda\beta/(1-\lambda)$, as stated in (10).

A.3 Projection Existence and Uniqueness: Proof of Proposition 2

We prove that (11) has a unique finite minimizer under the stated conditions.

Let $F_s(\rho) = T_s^{-1}\sum_t A_{h_t}(\rho)$ and $m_s = T_s^{-1}\sum_t \mathbb{E}_{\pi_\theta(\cdot|h_t)}[\Delta(\cdot \mid h_t)]$. Then $\Phi_s(\rho) = C_s + F_s(\rho) - \rho m_s$, giving

$$\Phi_s'(\rho) = F_s'(\rho) - m_s = \frac{1}{T_s}\sum_t \left(\mathbb{E}_{\pi_\rho(\cdot|h_t)}\Delta(\cdot \mid h_t) - \mathbb{E}_{\pi_\theta(\cdot|h_t)}\Delta(\cdot \mid h_t)\right), \tag{19}$$

$$\Phi_s''(\rho) = F_s''(\rho) = \frac{1}{T_s}\sum_t \mathrm{Var}_{\pi_\rho(\cdot|h_t)}(\Delta(\cdot \mid h_t)) \geq 0. \tag{20}$$

Under strictly positive teacher support and at least one prefix with nonconstant $\Delta(\cdot \mid h)$, $F_s''(\rho) > 0$ at every finite ρ, so F_s' is strictly increasing and continuous on $\mathbb{R}$.

Define $m_- = T_s^{-1}\sum_t \min_v \Delta(v \mid h_t)$ and $m_+ = T_s^{-1}\sum_t \max_v \Delta(v \mid h_t)$. As $\rho \to -\infty$, mass concentrates on $\operatorname{argmin}_v \Delta(v \mid h_t)$ for each t, so $F_s'(\rho) \to m_-$; as $\rho \to +\infty$, $F_s'(\rho) \to m_+$. When $m_s \in (m_-, m_+)$, the intermediate value theorem and strict monotonicity yield a unique finite root $\widehat{\rho}_s$ of $F_s'(\rho) = m_s$, which is also the unique minimizer of the strictly convex Φ_s. The moment-matching condition (12) is precisely $F_s'(\widehat{\rho}_s) = m_s$.

A strictly positive student has $m_s \in (m_-, m_+)$ whenever the student's per-prefix moments are in the open interior, which holds generically. Boundary moments ($m_s = m_-$ or $m_s = m_+$) would require the student to assign all mass to the token achieving the extremal contrast at every prefix, which in a language-model setting requires $\widehat{\rho}_s = \pm\infty$.

A.4 Residual Decomposition

We derive the exact decomposition of the student–target distance.

At a finite interior projection $\widehat{\rho}_s$, moment matching gives $F_s'(\widehat{\rho}_s) = m_s$. For any $r \in \mathbb{R}$,

$$\begin{aligned}\Phi_s(r) &= C_s + F_s(r) - r m_s \\ &= C_s + F_s(\widehat{\rho}_s) + F_s'(\widehat{\rho}_s)(r - \widehat{\rho}_s) + D_{F_s}(r, \widehat{\rho}_s) - r m_s \\ &= \Phi_s(\widehat{\rho}_s) + D_{F_s}(r, \widehat{\rho}_s),\end{aligned} \tag{21}$$

where $D_{F_s}(a,b) = F_s(a) - F_s(b) - F_s'(b)(a-b) \geq 0$ is the Bregman divergence of F_s. Expanding:

$$\boxed{\Phi_s(r) = R_s + D_{F_s}(r, \widehat{\rho}_s) = R_s + \frac{1}{T_s}\sum_t \mathrm{KL}(\pi_{\widehat{\rho}_s}(\cdot \mid h_t) \,\|\, \pi_r(\cdot \mid h_t))\,,} \tag{22}$$

where the second equality uses the identity $D_{F_s}(r, \widehat{\rho}_s) = T_s^{-1}\sum_t \mathrm{KL}(\pi_{\widehat{\rho}_s}(\cdot \mid h_t) \| \pi_r(\cdot \mid h_t))$, which follows by expanding the path KL in terms of A_{h_t} and Δ.

This shows that the student–target KL separates into an axis-fitting residual R_s and a distance from the projected teacher to the target. The signed coordinate identifies which target lies ahead of the student's projection; the residual R_s measures the mismatch between the student and the teacher path that the coordinate alone cannot capture.

A.5 Local Distance Approximation

Using (22) and the definition $I_s = \Phi_s''(\widehat{\rho}_s) = F_s''(\widehat{\rho}_s)$, a second-order Taylor expansion around $\widehat{\rho}_s$ gives

$$\Phi_s(\widehat{\rho}_s + \delta) = R_s + \tfrac{1}{2} I_s \delta^2 + O(\delta^3). \tag{23}$$

For small δ, the off-axis residual and local curvature thus approximate the student–target KL to second order. For arbitrary finite increments, the exact gap is given by the Bregman decomposition in (22).

A.6 Parameterization Equivalence

We derive the equivalence (25) in detail.

At a fixed prefix h, localization uses a detached snapshot of the current student to compute $\widehat{\rho}_s$ and β_s. During the subsequent update, the prefix, teachers, and these control quantities are held fixed; the student distribution $\pi_\theta(\cdot \mid h)$ in the loss remains differentiable. With conditioning on h suppressed:

$$\begin{aligned}
&\mathrm{KL}\big(\pi_\theta \| \pi^+\big) - \lambda\,\mathrm{KL}\Big(\pi_\theta \| \widetilde{\pi}^-_{\beta_s}\Big) \\
&\quad = \mathbb{E}_{\pi_\theta}\big[\log \pi_\theta - \log \pi^+\big] - \lambda \mathbb{E}_{\pi_\theta}[\log \pi_\theta - \log \pi_{-\beta_s}] \\
&\quad = (1-\lambda)\mathbb{E}_{\pi_\theta}[\log \pi_\theta] - \mathbb{E}_{\pi_\theta}\big[\log \pi^+ - \lambda \log \pi_{-\beta_s}\big]\,. \qquad (24)
\end{aligned}$$

Substituting $\log \pi_{-\beta_s} = \log \pi^+ - \beta_s \Delta - A_h(-\beta_s)$ and $\beta_s = (1-\lambda)/\lambda \cdot \rho_s^{\mathrm{target}}$:

$$\begin{aligned}
&= (1-\lambda)\mathbb{E}_{\pi_\theta}[\log \pi_\theta] - \mathbb{E}_{\pi_\theta}\big[\log \pi^+ - \lambda(\log \pi^+ - \beta_s \Delta)\big] - \lambda A_h(-\beta_s) \\
&= (1-\lambda)\mathbb{E}_{\pi_\theta}[\log \pi_\theta] - (1-\lambda)\mathbb{E}_{\pi_\theta}[\log \pi^+] - \lambda \beta_s \mathbb{E}_{\pi_\theta}[\Delta] + \text{const} \\
&= (1-\lambda)\left(\mathbb{E}_{\pi_\theta}[\log \pi_\theta] - \mathbb{E}_{\pi_\theta}\big[\log \pi^+ + \rho_s^{\mathrm{target}} \Delta - A_h(\rho_s^{\mathrm{target}})\big]\right) + \text{const} \\
&= (1-\lambda)\,\mathrm{KL}\Big(\pi_\theta \| \pi_{\rho_s^{\mathrm{target}}}\Big) + C_{h,s}, \qquad (25)
\end{aligned}$$

where $C_{h,s} = -\lambda A_h(-\beta_s) - (1-\lambda) A_h(\rho_s^{\mathrm{target}})$ is independent of the updated θ. This confirms (25): at a fixed detached target, the two parameterizations are equivalent up to a positive rescaling $(1-\lambda)$ and a constant, so their gradients with respect to θ are identical up to that scale. Averaging over the fixed rollout-prefix weights preserves this identity, including when different rollouts select different β_s. This is a conditional, per-update identity: optimization trajectories additionally depend on scaling and the projections recomputed along each run. Rollout sampling and target selection are refreshed between updates, without differentiating through either operation. Together with Propositions 1 and 2, these results establish the target construction, student projection, and conditional per-update equivalence.

A.7 Teacher-Axis Gradient Specificity

COPD and SACOPD use a fixed pair of teacher-axis anchors. The negative anchor π^- defines the contrast direction $\Delta = \log \pi^+ - \log \pi^-$. We analyze the local gradient with a controlled negative $\widetilde{\pi}^-_\beta = \pi_{-\beta}$, holding the prefix and β fixed during differentiation as in Appendix A.6. Static COPD with the natural negative is the special case $\beta = 1$.

Although the pair is fixed, its conditional distributions and contrast direction can vary with the context $h = (y_{<t}, x)$. A useful negative must not be identical to π^+, for which $\Delta \equiv 0$ and SACOPD produces no contrast signal. At the same time, broad unrelated differences between the teachers make the contrast less semantically specific. The desired regime is a fixed negative whose context-dependent differences from π^+ are concentrated primarily on the capability being transferred.

For the token loss $\ell_t = \mathrm{KL}(\pi_\theta \| \pi^+) - \lambda\,\mathrm{KL}(\pi_\theta \| \pi_{-\beta})$, define $c_{\lambda,\beta} = 1 - \lambda + \lambda\beta$. Substituting $\log \pi_{-\beta} = \log \pi^+ - \beta\Delta - A_h(-\beta)$, the derivative with respect to $\pi_\theta(v \mid h)$ at $\pi_\theta = \pi^-$ is $-c_{\lambda,\beta}\Delta(v \mid h)$ up to a token-independent constant. The softmax Jacobian centers this term, giving

$$\nabla_z \ell_t\big|_{\pi_\theta = \pi^-} = -c_{\lambda,\beta}\, \pi^-(\cdot \mid h) \odot \big[\Delta(\cdot \mid h) - \overline{\Delta}(h)\big]\,, \qquad \overline{\Delta}(h) = \mathbb{E}_{v \sim \pi^-(\cdot|h)}[\Delta(v \mid h)]. \quad (26)$$

Its squared dual Fisher norm is

$$\Big\| \nabla_z \ell_t\big|_{\pi_\theta = \pi^-} \Big\|^2_{F^\dagger} = c^2_{\lambda,\beta}\, \mathrm{Var}_{v \sim \pi^-(\cdot|h)}[\Delta(v \mid h)]. \qquad (27)$$

Here $F = \mathrm{diag}(\pi^-) - \pi^-(\pi^-)^\top$ is the logit Fisher matrix, and $\|g\|^2_{F^\dagger} = g^\top F^\dagger g$. For $\beta = 1$, the repulsive KL has zero gradient at $\pi_\theta = \pi^-$ and $c_{\lambda,1} = 1$, so the initial gradient equals that of

reverse-KL OPD. For a general fixed $\beta \geq 0$, it is scaled by $c_{\lambda,\beta} > 0$, with unchanged direction. Thus the gradient depends on both this scale and the centered contrast: the context-level average $\mathbb{E}_{\pi^-}[\Delta]$ is removed by the centering operation. The scaled variance in (27) measures local gradient energy, whereas gradient specificity is a property of where the contrast is concentrated across tokens and contexts. Under the intended pair design, π^- and π^+ differ mainly along the target capability, so the contrast-weighted update is concentrated on that capability rather than on incidental teacher differences. The external-negative comparison in Section 5 examines this teacher-pair design against a frozen initial-student negative.

A.8 Teacher Pair Instantiations

The SACOPD objective treats the teacher pair as a free design variable. Any fixed π^+ and π^- that differ along a behavioral axis of interest define a valid contrast direction $\Delta = \log \pi^+ - \log \pi^-$, and Proposition 1 supplies the corresponding virtual teacher without modification. The semantics of the resulting student are determined by which axis the pair encodes. We describe three natural choices.

In this work, the endpoint teacher pair is specified offline and held fixed throughout training. The rho-conditioned controller selects target positions along the path induced by these endpoints. We choose π^- to preserve the positive teacher's general behavior while differing primarily along the capability axis of interest.

- **Pre-/post-intervention.** π^- and π^+ are the same model before and after a targeted post-training step (e.g., π_{SFT} and $\pi_{\mathrm{SFT+RL}}$). Δ encodes what the intervention added.
- **Capability-gap.** π^- and π^+ are two same-family checkpoints at different scales (e.g., 1.5B and 32B). Δ encodes what scale provides.
- **Quality-conditioned.** π^+ and π^- are obtained by separately fine-tuning the same base model on the chosen and rejected sides of a preference dataset. Δ encodes the attribute that distinguishes chosen from rejected responses.

Each pair defines a semantically distinct Δ, and the same loss, closed-form optimum, and gradient analysis apply under the stated support assumptions.

A.9 Sequence-Level Interpretation

The same attraction–repulsion construction can separately be applied to complete response distributions. At a fixed prompt x, define

$$\mathcal{J}_x(p) = \mathrm{KL}\big(p \,\big\|\, \pi^+(\cdot \mid x)\big) - \lambda\, \mathrm{KL}\big(p \,\big\|\, \pi^-(\cdot \mid x)\big)\,. \tag{28}$$

For $0 < \lambda < 1$, common strictly positive teacher support, and a finite normalizer, the sequence-level virtual teacher is

$$\pi^\star_{\mathrm{seq}}(y \mid x) = \frac{1}{Z_{\mathrm{seq}}(x)} \pi^+(y \mid x)^\alpha \pi^-(y \mid x)^{1-\alpha}, \qquad \alpha = \frac{1}{1-\lambda}. \tag{29}$$

Expanding the KLs yields

$$\mathcal{J}_x(p) = (1-\lambda)\, \mathrm{KL}\big(p \,\big\|\, \pi^\star_{\mathrm{seq}}(\cdot \mid x)\big) - (1-\lambda) \log Z_{\mathrm{seq}}(x). \tag{30}$$

Thus, when the KLs and expectations are finite, $\pi^\star_{\mathrm{seq}}$ is the unique unrestricted sequence-level minimizer. This result concerns a different optimization problem from the fixed-prefix objective (2).

Conditional targets. Write $w(v \mid h) = \pi^+(v \mid h)^\alpha \pi^-(v \mid h)^{1-\alpha}$ and let hv denote the prefix extended by token v. Define the suffix partition function

$$B(hv) = \sum_{z:\, hvz \text{ complete}} \prod_j w(z_j \mid hvz_{<j}), \qquad \pi^\star_{\mathrm{seq}}(v \mid h) = \frac{w(v \mid h) B(hv)}{\sum_u w(u \mid h) B(hu)}. \tag{31}$$

Complete responses include a terminal token or a fixed horizon; the suffix partition function is one at a terminal prefix. In contrast, Proposition 1 gives $\pi^\star_{\mathrm{local}}(v \mid h) = w(v \mid h) / \sum_u w(u \mid h)$. These conditionals agree at a prefix only when $B(hv)$ is constant across its supported next tokens. Consequently, independently normalizing at each prefix does not in general recover the conditionals of the globally normalized sequence target.

Objective values and update gradients. For fixed teachers, exact student-policy sampling, and a sum over all response positions, the KL chain rule gives

$$\mathcal{J}(\theta) := \mathbb{E}_{x\sim\mathcal{D}}\mathcal{J}_x(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)} \sum_t \ell_\theta(h_t) = \sum_h d_\theta(h)\ell_\theta(h), \tag{32}$$

where $\ell_\theta(h) = \ell(\pi_\theta(\cdot \mid h); h)$ is the two-term conditional loss in (15), and d_θ is the unnormalized prefix visitation measure. Under the usual conditions permitting differentiation under the expectation,

$$\nabla_\theta \mathcal{J}(\theta) = \sum_h d_\theta(h)\nabla_\theta \ell_\theta(h) + \sum_h \ell_\theta(h)\nabla_\theta d_\theta(h). \tag{33}$$

The fixed-prefix update differentiates the conditional loss with its sampling weights held fixed; it does not include the second term. Moreover, ν_k reflects the actual rollout temperature, response window, and loss aggregation, rather than assuming a complete-response sum. The algorithm is therefore defined by the per-update conditional objective and its local target identity, not by the full gradient of (28).

B Experimental Protocol and Implementation Details

B.1 Matched Training Configurations

Teacher provenance. We directly use the public positive-teacher checkpoints released by the authors of G-OPD (Yang et al., 2026):

- Mathematics: `Keven16/Qwen3-4B-Non-Thinking-RL-Math-Step500`.
- Code: `Keven16/Qwen3-4B-Non-Thinking-RL-Code-Step300`.

These experts were obtained by applying GRPO to Qwen3-4B Base for 500 steps on DeepMath-103K examples with difficulty level at least six (He et al., 2026) and 300 steps on Eurus-RL-Code (Cui et al., 2025), respectively. Both positive teachers remain frozen during distillation. The student starts from their shared pre-RL Qwen3-4B Base initialization, whose frozen copy serves as the negative teacher.

The mathematical and code experiments instantiate the same construction on task-specific teacher axes. Table B1 records the settings shared by the matched OPD, COPD, and SACOPD training runs. ExOPD and IW-OPD retain their own training budgets; the matched-budget W2S-OPD adaptations are specified separately in Appendix B.4.1. The response window in Table B1 defines the segment used for the training loss and SACOPD projection; evaluation uses the longer generation caps described below.

The mathematical pool is obtained by filtering DeepMath-103K to difficulty level six or above. Its 57,046 examples are split into 56,918 training prompts and 128 held-out audit prompts. The code route uses all 25,276 Eurus-RL-Code examples. SACOPD computes full-vocabulary student, positive-teacher, and anchor distributions on each observed prefix. The fitted projection and resolved contrast strength are detached before the model update. Audits are run on common fixed responses and checkpoint-own on-policy responses and record $\widehat{\rho}$, residual R_s, curvature I_s, numerical root status, target position, and contrast strength.

Hyperparameters and checkpoints. All trained methods in Table 1 use their final step-50 checkpoint. The main configurations use $\lambda = .80$ for mathematics and .20 for code, with SACOPD lookahead $\delta = 2$ and 1, respectively (Table B1). The code configuration is shared across all three suites. Section 5.2 examines sensitivity around the mathematical main configuration. ExOPD and IW-OPD coefficients follow their released recommendations (Appendix B.4.2); W2S-OPD includes both the recommended domain-specific coefficients and an additional $\alpha = 5$ setting (Appendix B.4.1).

B.2 Evaluation Profiles

All primary generation is non-thinking with temperature 1, top-p 1, and top-k -1. The mathematical main comparison uses generation seed 2, sample-major request order, and 120 problems with 32

Table B1: Main-training configurations for OPD, COPD, and SACOPD in mathematics and code. Optimization and rollout settings apply to all three methods.

Setting	Mathematics	Code
Student / anchor	Qwen3-4B Base	Qwen3-4B Base
Positive teacher	Math-RL step 500	Code-RL step 300
Training data	DeepMath-103K level ≥ 6	Eurus-RL-Code
Pool / actual train samples	57,046 / 56,918	25,276 / 25,276
OPD/COPD/SACOPD updates	50	50
Global batch / training seed	6 / 1	6 / 1
Optimizer / learning rate	AdamW / 5×10^{-6}	AdamW / 5×10^{-6}
Learning-rate schedule	Constant	Constant
AdamW (β_1, β_2) / ϵ	$(.9, .999)$ / 10^{-8}	$(.9, .999)$ / 10^{-8}
Weight decay	.01	.01
Training response window	256 tokens	256 tokens
Training temperature / top-p	$.6/1$	$.6/1$
Training top-k / thinking	Unrestricted / disabled	Unrestricted / disabled
COPD	$\lambda = .80,\ \beta = 1$	$\lambda = .20,\ \beta = 1$
SACOPD	$\lambda = .80,\ \delta = 2$	$\lambda = .20,\ \delta = 1$

responses each. Its response cap is 32,768 tokens and its model context is 40,960 tokens. The exact scorer reads the final complete `\boxed{}` expression from the full response and verifies it with Math-Verify.

Supporting experiments state their sample counts and generation caps in the corresponding experimental descriptions. Scores are reported separately for each value of K.

Parameter-sensitivity readouts. Table 2 reports parameter sensitivity at 32K/$K = 8$, with $\lambda = .80, \delta = 2$ as the main setting. The lookahead panel varies $\delta \in \{1, 1.5, 2, 2.5, 3\}$ at fixed $\lambda = .80$; the parameterization panel varies λ at fixed $\delta = 2$. Controller values average the six rollouts in the final on-policy batch; the checkpoint audit uses a shared 128-response replay. The two shaded rows reuse the same checkpoint and the same generated responses. The other parameterization rows are independently trained with the same seed and update budget.

B.3 Code Evaluation Protocol

HumanEval+ contains 164 problems and MBPP+ contains 378; both use four responses per problem, a 16,384-token response cap, and a 32,768-token model context. A sample is correct only when it passes both the EvalPlus base and plus tests. We replace the default EvalPlus sanitizer with a deterministic, entrypoint-aware extraction procedure, applied uniformly to all methods, while retaining the official base-plus correctness criterion. Specifically, we select the last fenced code block containing the function named in the task, or, if no such block exists, extract the unfenced function using tokenization and indentation. For an unfenced function, the selected source region also includes immediately preceding import statements at the same indentation, separated only by blank lines. The boundary stops at prose, fences, other definitions, or other statements. We preserve code inside the selected region, including example assertions and calls; imports are not collected from earlier drafts. Extraction uses only the response and the task's public function name, without test outcomes or reference solutions.

All methods use the same execution limits: a per-test timeout of at least 10 seconds or eight times the reference runtime, whichever is larger, 4 GiB of memory, and single-threaded numerical libraries.

Following G-OPD (Yang et al., 2026), we use the February–May 2025 subset of LiveCodeBench v6: 131 problems, comprising 49 functional and 82 standard-input/output tasks. This selects problems dated February 1–May 31, 2025 from the v6 increment, rather than the cumulative `release_v6` benchmark. For functional tasks, the prompt includes the question and the benchmark's public

starter code, specifying the required class or function name and signature. It also states that the evaluator calls this interface and expects a return value. Standard-input/output tasks use the question prompt. All methods use these same prompt conventions. Evaluation uses four responses per problem, a 32,768-token response cap, a 40,960-token context, and the official execution scorer.

B.4 External-Method Baselines

The first three rows of Table 1 report the Qwen3-4B Base initialization, the positive teacher, and matched OPD. The OPD, COPD, and SACOPD training runs share the student initialization, data order, optimizer, seed, and budget (Table B1).

B.4.1 Matched-Budget W2S-OPD Adaptations

We adapt the W2S-OPD objective (Yu et al., 2026) to the task-specific 4B axes and training pools in Table B1. The student, frozen student-base anchor, and negative teacher are Qwen3-4B Base; the positive teacher is Math-RL step 500 or Code-RL step 300. This is a matched-budget, same-size adaptation of the objective, rather than a reproduction of the original 8B weak-to-strong setting with batch 64, 100 updates, and a 16K training response window.

The proxy teacher has logits $z_{\text{base}} + \alpha(z_+ - z_-)$, and training minimizes reverse KL on the proxy teacher's top 32 tokens without additional loss rescaling. The first setting retains the original paper's domain-specific coefficients. Since our anchor and negative coincide, the untruncated proxy distribution is

$$q_\alpha(v \mid h) \propto \pi^-(v \mid h) \left(\frac{\pi^+(v \mid h)}{\pi^-(v \mid h)} \right)^\alpha. \tag{34}$$

Thus $\alpha = 1$ recovers the positive teacher, giving top-32 reverse-KL OPD, while $\alpha = .75$ interpolates between Base and the positive teacher. Neither setting extrapolates beyond the positive teacher on this matched axis.

We additionally evaluate $\alpha = 5$ to match the target strength of the main mathematical COPD configuration: with the natural negative ($\beta = 1$), Eq. (3) gives $\alpha = 1/(1-\lambda) = 5$ at $\lambda = .8$. This aligns the untruncated virtual target; the implemented losses retain their respective vocabulary support and scaling. W2S-OPD uses top-32 reverse KL with scale one, whereas COPD uses the full-vocabulary contrastive loss, whose equivalent target-KL coefficient is $1 - \lambda = .2$ for this math setting. The two runs therefore compare implementations of the same fixed target rather than identical optimization procedures. For code, we carry over $\alpha = 5$ as a stronger fixed-target baseline; the main code COPD setting $\lambda = .2$, $\beta = 1$ instead corresponds to $\alpha = 1.25$. Both W2S-OPD settings retain a fixed coefficient without student-state conditioning.

Table B2: W2S-OPD configurations in Table 1. Each setting has separate math and code checkpoints. Shared settings apply to both domains.

Setting	W2S-OPD	W2S-OPD ($\alpha = 5$)
α (math / code)	$1.0/.75$	$5.0/5.0$
Proxy top-k / loss scale	$32/1$	$32/1$
Updates / global batch	$50/6$	$50/6$
Rollouts / response cap	$300/256$	$300/256$
Optimizer / learning rate	AdamW / 5×10^{-6}	AdamW / 5×10^{-6}
Training seed	1	1
Training temperature / top-p	$.6/1$	$.6/1$
Evaluated checkpoint	step 50	step 50

Training disables thinking and uses ZeRO-2 with gradient checkpointing. The code prompt window is 2048 tokens; mathematics uses the full prompt.

Both settings follow the main-table benchmark and scoring protocol: generation seed 2, temperature/top-p $1/1$, unrestricted top-k, non-thinking generation, and sample-major request order. Math uses 32 samples per problem, a 32K response cap and 40K context, with the final-box

Math-Verify scorer. Code uses four samples and the caps, prompts, extraction, and official scorers in Appendix B.3. Each run has complete coverage: 3840 math, 656 HumanEval+, 1512 MBPP+, and 524 LiveCodeBench responses.

The stronger-target setting also changes generation cost: average math response length is 18,067 versus 11,637 tokens, with 32K cap-hit rates of 8.33% versus $.10\%$. HumanEval+/MBPP+/LiveCodeBench average lengths are 7295/4946/19,783 tokens for $\alpha = 5$ and 752/562/2728 for the domain-specific setting.

B.4.2 External Reproductions: ExOPD and IW-OPD

ExOPD and IW-OPD are external objectives re-run inside the same student–teacher axis. **ExOPD** (Yang et al., 2026) is the reward-extrapolation regime of Generalized OPD, which scales the teacher–reference log-ratio reward by a coefficient $\lambda_{\text{vals}} > 1$; we use the released single-teacher branch at its recommended $\lambda_{\text{vals}} = 1.25$, with the reference set to the student's base initialization. **IW-OPD** (Xie et al., 2026) reweights token-level OPD advantages by the cumulative teacher–student disagreement along the sampled response; we use the released implementation with $\lambda_{\text{vals}} = 1.0$, a maximum weight of 1.5, and the absolute-disagreement variant that the release enables by default.

Both external methods use the same student, teacher anchors, and training pool as the matched methods: Qwen3-4B Base, the Math-RL step-500 or Code-RL step-300 teacher, and the DeepMath-103K (level ≥ 6) or Eurus-RL-Code pool. This adapts the released model pairings (a smaller student for G-OPD and a 30B teacher for IW-OPD) to the task-specific Qwen3-4B axes studied here. Both panels use AdamW at lr 10^{-5} with no warmup, one rollout per prompt, temperature and top-p of 1, thinking disabled, seed 42, token-level rollout correction, and remove-padding attention. The matched-budget panel uses a 256-token response cap; the released-scale panel uses the 16,384-token cap of the earlier reproduction.

The main-table external panel uses global batch 6, 50 updates, and a 256-token training response cap, matching the $6 \times 50 = 300$ rollouts used by OPD, COPD, and SACOPD. The matched-budget runs use the same optimizer, seed, teacher axes, and evaluation protocol as the released-scale runs described below.

For comparison with the released training scale, the released-scale reproductions use global batch 126 and 50 updates, giving $126 \times 50 = 6300$ training rollouts. The released recipes use batch 1024; the same-size ExOPD schedule also uses 50 updates, while the released IW-OPD schedule uses three epochs (≈ 167 updates at its original batch size). These runs use a 16,384-token training response window. Their results are retained in Table B3; the main table reports the matched-budget counterparts.

We use $\lambda_{\text{vals}} = 1.25$ as the main external reference because it is the setting recommended by the original work; the non-extrapolative coefficient remains documented in the source evaluation record.

Table B4 records the released-scale configurations, and Table 1 gives the complete matched-budget per-source and per-suite results. Both methods are evaluated at their final checkpoints on the same 120-problem mathematics panel and code suites used in Table 1, with the same sample counts, generation caps, and official scorers.

B.5 Teacher-Pair and Target-Selection Protocols

Tables 3 and 4 use a 4B student, 50 updates, global batch 6, training seed 1, a 32K response cap, and GOPD-exact scoring. The teacher-pair panel uses 120×8 responses. Pre-/post-RL and cross-specialist pairs share the Math-RL OPD reference (sample indices 0–7 of its $K = 32$ evaluation); dense-size uses its own matched OPD. Pre-/post-RL COPD has 572/960 correct responses and 89/120 covered problems. Reported gains use raw counts before rounding.

B.5.1 Additional Teacher-Pair Results

Table B5 complements the 4B-student comparison in Table 3. With a 1.7B student, both the pre-/post-RL and cross-specialist pairs yield substantial gains over their corresponding OPD references: mean@8 increases by 21.04 and 20.00 percentage points, respectively. These results extend the observed benefits of the two teacher contrasts to a smaller student. For the same 1.7B student,

Table B3: Released-scale external reproductions retained from the earlier panel. Entries are mean@K / pass@K (%). These runs use 6300 training rollouts; the main table reports the matched-budget counterparts.

Method	AIME24		AIME25		HMMT-F		HMMT-N		Math Overall	
	m@32	p@32	m@32	p@32	m@32	p@32	m@32	p@32	m@32	p@32
IW-OPD	64.38	83.33	58.75	80.00	35.10	60.00	37.29	70.00	48.88	73.33
ExOPD	67.29	80.00	63.44	80.00	35.31	56.67	41.77	66.67	51.95	70.83

Method	HE+		MBPP+		LCB		Code Overall	
	m@4	p@4	m@4	p@4	m@4	p@4	m@4	p@4
IW-OPD	88.41	94.51	72.95	81.75	38.74	49.62	66.70	75.29
ExOPD	91.01	96.95	73.15	80.69	44.85	48.85	69.67	75.50

Table B4: Configurations of the ExOPD and IW-OPD reproductions. Official values are in parentheses.

Method	Codebase	Objective setting	Updates	Batch
ExOPD (both domains)	G-OPD	single-teacher, $\lambda_{\text{vals}} = 1.25$	50 (50)	126 (1024)
IW-OPD (both domains)	IW-OPD	$\lambda_{\text{vals}} = 1.0$, $w_{\max} = 1.5$, abs	50 ($\approx$167)	126 (1024)

the dense-size configuration at $\lambda = .20$ reaches 9.58% mean@8 and 22.50% pass@8, recovering performance near its OPD reference and illustrating the role of axis-specific strength calibration.

Table B5: Additional teacher-pair results with a 1.7B student and static COPD ($\beta = 1$). OPD and COPD entries are mean@8 / pass@8 (%); gains are percentage points relative to the corresponding OPD. Math-RL, Code-RL, and Base teachers are 4B unless specified. The final row uses the post-hoc lower-strength setting $\lambda = .20$.

Positive / negative	λ	OPD	COPD	Δmean / Δpass
Math-RL / Base	.80	15.10 / 31.67	36.15 / 58.33	+21.04 / +26.67
Math-RL / Code-RL	.80	16.04 / 35.83	36.04 / 55.83	+20.00 / +20.00
14B / Base	.20	8.96 / 21.67	9.58 / 22.50	+0.63 / +0.83

All three rows use 50 training updates, global batch 6, training seed 1, and a 256-token training response window, with thinking disabled. Evaluation uses the same 120-problem mathematics panel, $K = 8$, a 32K response cap, and GOPD-exact scoring. The pre-/post-RL row uses Lite-OPD's embedded evaluation runtime; the cross-specialist and dense-size rows use vLLM. Each row uses an OPD reference evaluated under its corresponding protocol. All three COPD runs use the fixed-β controlled training implementation.

B.6 Reproducibility across Training Seeds

We evaluate OPD, COPD, and SACOPD across training seeds 1, 2, and 3 in both domains. Both contrastive methods improve overall accuracy and coverage over OPD in each seed (Table B6). Averaged over the three seeds, COPD and SACOPD improve mathematical mean@32 by 12.59 and 13.02 percentage points, respectively; their code mean@4 gains are 2.40 and 4.16 points. The repeated runs preserve the large mathematical gain from static extrapolation and the additional code benefit from state-aware target selection.

These runs retain each method's teacher pair, training-data recipe, hyperparameters, 50-update budget, global batch of 6, and 256-token training response window. Generation seed is fixed at 2 in both

Table B6: Results across three training seeds for OPD, COPD, and SACOPD. All scores are percentages. Math and Code denote the same within-domain macro-averages as Table 1. The final block reports mean ± sample standard deviation over seeds 1–3 ($n = 3$); shading highlights the contrastive methods in this summary.

		Math ($K = 32$)		Code ($K = 4$)	
Training seed	Method	mean@32	pass@32	mean@4	pass@4
1	OPD	46.17	73.33	66.79	75.95
	COPD	58.91	79.17	67.98	76.93
	SACOPD	59.48	80.83	69.77	78.30
2	OPD	47.71	74.17	66.14	75.53
	COPD	59.06	80.83	69.16	76.52
	SACOPD	59.43	80.00	70.39	77.82
3	OPD	45.78	74.17	64.94	75.56
	COPD	59.45	80.00	67.94	76.61
	SACOPD	59.82	81.67	70.19	77.79
Seeds 1–3	OPD	46.55 ± 1.02	73.89 ± 0.48	65.96 ± 0.94	75.68 ± 0.24
	COPD	59.14 ± 0.28	80.00 ± 0.83	68.36 ± 0.70	76.69 ± 0.22
	SACOPD	59.57 ± 0.21	80.83 ± 0.83	70.11 ± 0.32	77.97 ± 0.29

domains, including the original three methods. Table B7 gives the corresponding benchmark-level results.

Table B7: Benchmark-level results for the additional training seeds. Each entry is mean@K / pass@K (%), with $K = 32$ for mathematics and $K = 4$ for code.

(a) Mathematics ($K = 32$)

Seed	Method	AIME 2024	AIME 2025	HMMT Feb.	HMMT Nov.
2	OPD	61.04 / 83.33	56.04 / 83.33	33.33 / 63.33	40.42 / 66.67
	COPD	74.90 / 90.00	68.85 / 83.33	41.98 / 73.33	50.52 / 76.67
	SACOPD	73.96 / 90.00	68.96 / 86.67	42.71 / 70.00	52.08 / 73.33
3	OPD	59.06 / 83.33	54.58 / 83.33	32.60 / 70.00	36.88 / 60.00
	COPD	73.02 / 90.00	68.75 / 83.33	43.54 / 73.33	52.50 / 73.33
	SACOPD	73.54 / 90.00	71.46 / 86.67	42.40 / 70.00	51.88 / 80.00

(b) Code ($K = 4$)

Seed	Method	HumanEval+	MBPP+	LiveCodeBench
2	OPD	85.82 / 93.90	73.08 / 83.07	39.50 / 49.62
	COPD	89.94 / 95.12	73.28 / 82.54	44.27 / 51.91
	SACOPD	89.94 / 93.90	75.99 / 83.07	45.23 / 56.49
3	OPD	84.45 / 94.51	71.63 / 82.54	38.74 / 49.62
	COPD	88.26 / 95.12	72.22 / 82.80	43.32 / 51.91
	SACOPD	90.09 / 93.29	76.19 / 83.60	44.27 / 56.49

C Complete Static Target-Position Sweep

Table C8 gives the exact values underlying Figure 1. The sweep varies the fixed virtual-target position, not the rho-conditioned lookahead. All rows use $\lambda = .80$, training seed 1, and 50 updates, evaluated on 120 mathematics problems under the same 32K/$K = 8$ exact protocol. The highest

observed mean@8 and pass@8 occur at $\beta = 1.10$; this setting has not been evaluated at $K = 32$. The curve identifies an effective region within the target family available to SACOPD, rather than predicting the outcome of adaptive training. SACOPD resolves β from the student coordinate and fixed lookahead on each rollout, so it does not require selecting a single fixed strength from this grid.

Table C8: Static target-position sweep at 32K/$K = 8$. Accuracy and coverage are percentages; bold marks the highest observed value within this sweep.

Fixed β	Target ρ_β	mean@8	pass@8	Correct	Covered
.25	1.0	56.15	71.67	539/960	86/120
.50	2.0	57.81	71.67	555/960	86/120
.75	3.0	57.92	70.83	556/960	85/120
.90	3.6	58.54	74.17	562/960	89/120
1.00	4.0	59.58	74.17	572/960	89/120
1.10	4.4	**60.00**	**75.00**	576/960	90/120
1.25	5.0	59.69	72.50	573/960	87/120
1.50	6.0	57.92	72.50	556/960	87/120

D Coordinate Diagnostics and Training Dynamics

D.1 Common-Replay Protocol

The positive teacher is Qwen3-4B Math-RL step 500 and the anchor is Qwen3-4B Base. All coordinate measurements in this section use the same 128 development responses, one per held-out prompt, and the first 256 response prediction prefixes. For each checkpoint and response, we evaluate full-vocabulary conditional distributions and solve the reverse-KL projection in Eq. (11). The checkpoint readout is the mean of the 128 response-level fitted coordinates. Keeping the replay fixed isolates policy changes on common observed states. Projection uses the student and teacher conditional distributions at these prefixes.

The longitudinal panel follows one static COPD run with $\lambda = .80$, the natural Base negative, training seed 1, and global batch 6. Its checkpoints are observed at steps 0, 10, 20, 25, 40, 50, 75, 100, and 150; step 0 is the Base initialization. The positive teacher is audited separately as a calibration anchor. The audit uses bfloat16 model forwards, the same solver settings, and the same observation window for all policies. All 10×128 fitted projections are valid.

D.2 Anchor Recovery and Checkpoint Localization

Table D9 reports both location and projection quality. Base and positive exactly recover coordinates -1 and 0, with mean residuals of 1.29×10^{-9} and 1.69×10^{-9}, respectively. At step 10 every replay coordinate is negative; by step 20 every coordinate is positive. Thus an observed crossing occurs between these two checkpoints. Subsequent mean coordinates continue to advance through step 150. Residual R_s and curvature I_s accompany the coordinate to distinguish axis position, off-axis discrepancy, and local identification.

Across student checkpoints from step 10 to 150, mean residual ranges from 0.0275 to 0.0825 and mean curvature from 0.0638 to 0.1625. After the initial crossing at step 20, mean residual decreases overall from 0.0825 to 0.0566 at step 150, while mean coordinate continues to advance. Every response-level curvature is positive, with a minimum of 0.0219 at step 150. Together, these measurements track axis position, projection fit, and local sensitivity throughout training.

D.3 Population Shift and Paired Coordinate Movement

Figure 2b displays the empirical CDFs at steps 10, 25, and 50. The same responses are evaluated at every checkpoint, so we can also compare each response with itself over training. Table D10 covers all seven adjacent observed intervals from step 10 through step 150. Every paired change is positive

Table D9: Common-replay localization. R and I are mean residual and curvature; P10–P90 is the 10th–90th percentile range of response coordinates.

Policy	Mean $\widehat{\rho}$	Median	P10–P90	R	I	$\widehat{\rho} > 0$
Base / step 0	-1.0000	-1.0000	$[-1.000, -1.000]$	1.29×10^{-9}	0.7583	0/128
Positive teacher	0.0000	0.0000	$[0.000, 0.000]$	1.69×10^{-9}	0.1246	0/128
Step 10	-0.2844	-0.2959	$[-0.381, -0.163]$	0.0275	0.1625	0/128
Step 20	0.7842	0.7933	$[0.559, 1.053]$	0.0825	0.1283	128/128
Step 25	0.9383	0.9534	$[0.678, 1.237]$	0.0808	0.1262	128/128
Step 40	1.1812	1.1716	$[0.884, 1.482]$	0.0713	0.1103	128/128
Step 50	1.2866	1.2651	$[0.984, 1.614]$	0.0657	0.1004	128/128
Step 75	1.4514	1.4158	$[1.147, 1.746]$	0.0578	0.0866	128/128
Step 100	1.5602	1.5708	$[1.239, 1.850]$	0.0553	0.0781	128/128
Step 150	1.7597	1.7745	$[1.435, 2.050]$	0.0566	0.0638	128/128

in every interval. This establishes a population-wide progression on the observed replay, beyond an increase in its mean or upper tail.

Confidence intervals use 5,000 bootstrap resamples over the 128 replay prompts, with bootstrap seed 20260920 and the checkpoint pairing retained in each resample. The intervals quantify variation across replay prompts within this single training run. For the two intervals highlighted by the CDF panel, the mean changes are 1.2227 for step 10 to 25 (95% CI [1.1852, 1.2607]) and 0.3483 for step 25 to 50 ([0.3200, 0.3780]).

Table D10: Paired coordinate changes between checkpoints on the same replay responses.

Steps	Mean $\Delta\widehat{\rho}$	95% bootstrap CI	Positive changes
$10 \to 20$	1.0686	[1.0343, 1.1034]	128/128
$20 \to 25$	0.1541	[0.1405, 0.1680]	128/128
$25 \to 40$	0.2429	[0.2214, 0.2649]	128/128
$40 \to 50$	0.1054	[0.0960, 0.1153]	128/128
$50 \to 75$	0.1648	[0.1515, 0.1790]	128/128
$75 \to 100$	0.1088	[0.0994, 0.1187]	128/128
$100 \to 150$	0.1996	[0.1845, 0.2158]	128/128

D.4 Coordinates and Downstream Performance

We pair the coordinate measurements with the same run's original training-exposure evaluations at steps 20, 40, 50, 75, 100, and 150. Each evaluation uses the four-source 120-problem mathematics set, four responses per problem, a 32,768-token cap, and GOPD-exact scoring. All six accuracy points use the same Lite-OPD built-in inference runtime for generation; the main-table evaluations use vLLM. The downstream evaluation set and the development replay serve different measurement roles; they are paired by checkpoint rather than by problem.

Table D11 and Figure 3 show that mean coordinate advances from 0.7842 to 1.7597 while mean@4 stays within 57.71–60.21%. This makes continued policy motion visible within an accuracy plateau. The descriptive Pearson association across these six time-ordered points is 0.2773. The coordinate therefore supplies an optimization-state readout, while task performance measures the realized benefit of that motion.

Accuracy bands retain the original exposure panel's 5,000 problem-bootstrap resamples of the 120 evaluation problems, retaining each problem's four samples. Sample-level coverage and exact-correct counts were checked against each original aggregate. The $K = 4$ evaluations here describe training dynamics; their responses are kept separate from the main-table $K = 32$ evaluations and the cross-sectional $K = 8$ panel below.

Table D11: Coordinate and performance across checkpoints. Accuracy and cap-hit rates are percentages; response length is in thousands of tokens.

Step	Mean $\widehat{\rho}$	mean@4	pass@4	Tokens (K)	Cap hit
20	0.7842	59.17	72.50	15.05	2.71
40	1.1812	57.71	68.33	17.01	6.46
50	1.2866	59.79	70.00	17.24	5.62
75	1.4514	60.21	70.83	18.13	8.12
100	1.5602	58.54	70.83	18.82	9.79
150	1.7597	59.79	69.17	19.16	8.96

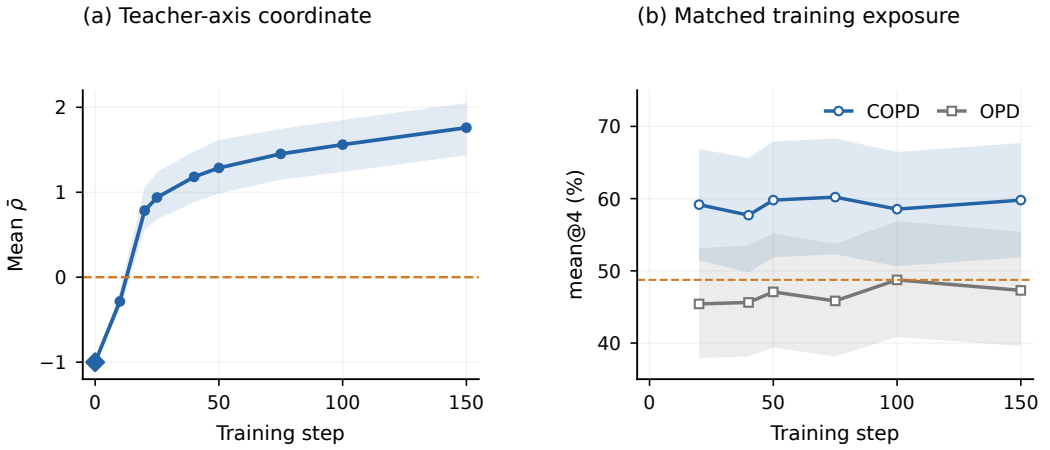


Figure 3: Coordinate evolution and matched training exposure. **(a)** Mean static COPD coordinate; shading spans the 10th–90th percentiles across the fixed 128 responses, and the dashed line marks the positive teacher at zero. **(b)** COPD and OPD mean@4 with 95% problem-bootstrap intervals; the dashed line marks the best OPD score (48.75%). Each training step uses six on-policy rollouts.

D.5 Cross-Configuration Association and Panel Sensitivity

The cross-sectional panel contains 10 unique step-50 checkpoints: five fixed-β configurations ($\beta \in \{0.25, 0.50, 0.75, 0.90, 1.00\}$) and five continuous-$\delta$ configurations ($\delta \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$). These checkpoints come from training runs separate from the longitudinal trajectory and are audited on the same teacher axis and fixed 128-response replay as that trajectory. All have matched $K =$ 8/32K GOPD-exact evaluations with 960 responses per checkpoint. Table D12 lists all 10 points shown in Figure 4. The panel compares configurations under one training seed.

The ten-point panel has Pearson $r = 0.8110$ and Spearman $r_s = 0.4073$ for mean@8. Within the fixed-β family, coordinate and mean accuracy are positively associated ($r = 0.8959$); within the continuous-δ family, positions concentrate in the higher-coordinate region and the association is negative ($r = -0.8624$). Together with the longitudinal plateau, these patterns support interpreting the coordinate as an axis-relative optimization state, with its mean-accuracy association varying across configuration families.

D.6 Training-Exposure Comparison

The contrastive gain is established after 120 training rollouts. Figure 3(b) sweeps training exposure from 20 to 150 updates (120 to 900 on-policy rollouts) for COPD and matched OPD under the same training recipe, evaluated at $K = 4$ per checkpoint. COPD reaches 59.17% mean@4 after 120 rollouts and stays within 57.71–60.21% through 900 rollouts. Matched OPD rises from 45.42% to 47.29% over the same sweep; its best reading anywhere in the sweep, 48.75% at 600

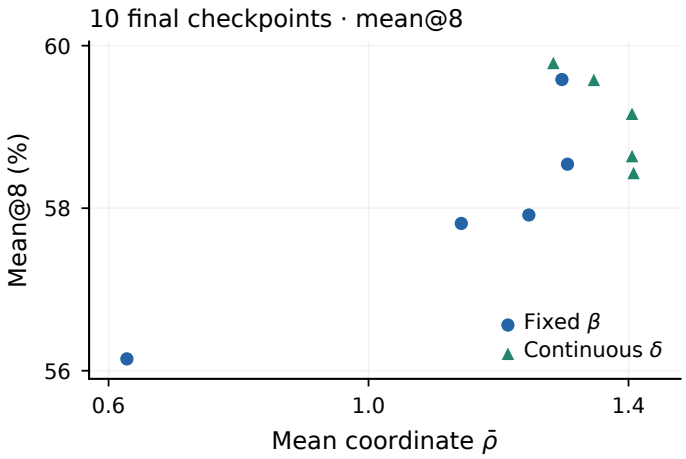


Figure 4: Mean coordinate and mean@8 across 10 final checkpoints. Colors and symbols distinguish configuration families.

Table D12: Mean coordinate and mean@8 across configurations.

Configuration	Mean $\widehat{\rho}$	mean@8 (%)
Fixed $\beta = 0.25$	0.6282	56.15
Fixed $\beta = 0.50$	1.1427	57.81
Fixed $\beta = 0.75$	1.2466	57.92
Fixed $\beta = 0.90$	1.3061	58.54
Fixed $\beta = 1.00$	1.2975	59.58
Continuous $\delta = 1.0$	1.2843	59.79
Continuous $\delta = 1.5$	1.3467	59.58
Continuous $\delta = 2.0$	1.4054	59.17
Continuous $\delta = 2.5$	1.4055	58.65
Continuous $\delta = 3.0$	1.4078	58.44

rollouts, remains below COPD at every matched checkpoint. The paired difference is positive at all six checkpoints, ranging from $+9.79$ to $+14.38$ pp, with problem-level bootstrap intervals excluding zero in every case.

E Qualitative Case Studies

We examine two selected problems from the evaluations in Table 1. The figures reproduce saved responses from OPD, COPD, and SACOPD at the same sample index (2), preserving their wording and source order; bracketed ellipses mark omissions.

E.1 Mathematical Reasoning

The HMMT polynomial problem distinguishes a common annihilating polynomial from a minimum-degree polynomial taking equal values at two specified arguments (Figure 5). SACOPD directly tests a cubic, derives its coefficients, and verifies equality at both arguments, obtaining the correct answer of 890. The displayed OPD response instead identifies the required polynomial with the quartic minimal polynomial; COPD attempts a cubic but rejects it after an incorrect expansion. Across all 32 samples of this problem, OPD, COPD, and SACOPD answer correctly 16, 26, and

30 times, respectively. The displayed SACOPD response is also shorter than the COPD response, illustrating the importance of the reasoning path beyond response length.

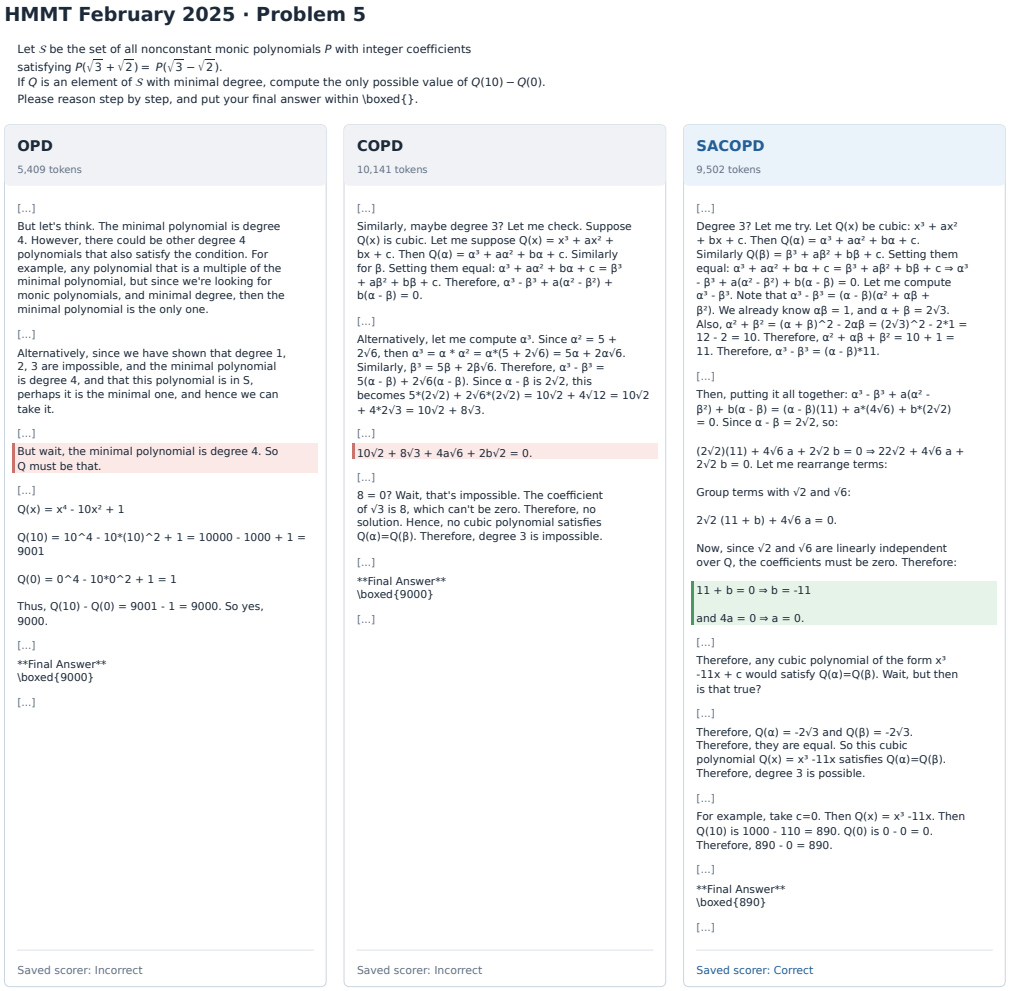


Figure 5: Original response excerpts for HMMT February 2025, problem 5. Red highlights the incorrect inference or calculation; green highlights the coefficients yielding a valid cubic. Token counts refer to the complete responses.

E.2 Algorithm Design

On AtCoder ABC400 E, SACOPD recognizes that each valid number is a perfect square, reducing the search domain from numbers up to 10^{12} to their roots up to 10^6 (Figure 6). Its final program counts distinct prime factors with a sieve, squares the eligible values, and answers queries by binary search. This connects a structural observation in the response to an executable algorithm that passes the tests. The displayed OPD and COPD programs enumerate all distinct prime pairs from their prime lists and exceed the execution time limit. Across four samples of this problem, OPD, COPD, and SACOPD pass 0, 0, and 4 times, respectively; the displayed sample uses the square-root sieve. These qualitative examples complement the aggregate results by showing how SACOPD responses construct a valid polynomial and turn a structural observation into an efficient algorithm.

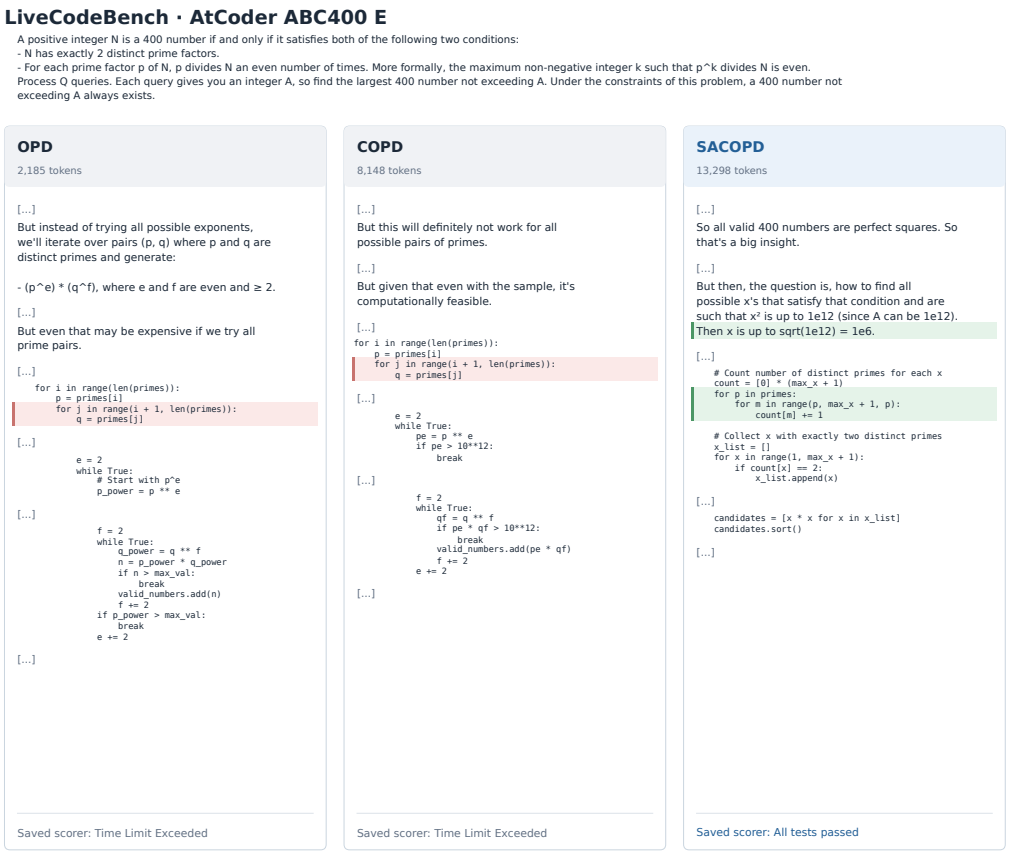


Figure 6: Original response and submitted-code excerpts for AtCoder ABC400 E. Red highlights the unrestricted prime-pair loops; green highlights the reduced search domain and the corresponding sieve. Results are from the saved official execution scorer.