

Composing What Each Teacher Learned: Multi-Teacher On-Policy Distillation through Teacher-Relative Shifts

Hejian Sang^{1*} Zhengze Zhou^{2*} Shayan Mohajer Hamidi
Xiaomin Li³ Rohit Jain² Alborz Geraimifard²

¹Iowa State University ²LinkedIn ³Harvard University

Abstract

Multi-teacher on-policy distillation (MOPD) is used in two settings. In *common-domain composition*, several teachers score each student rollout from one prompt domain and their signals form a single target; in *routed-domain distillation*, prompts from different domains are assigned to the corresponding specialist. Both settings usually transfer each teacher’s endpoint policy, which mixes what post-training changed with preferences inherited from the teacher’s base. We introduce Δ -MOPD, which transfers each teacher’s teacher-minus-base logit shift re-anchored at the student’s frozen initialization, and compare it with endpoint supervision in both settings while holding teacher selection fixed. We first expose the mechanism that impedes endpoint transfer: inherited base pull can exceed the post-training shift. Removing it reduces the teacher-term norm ratio from 5.2 : 1 to 1.44 : 1 and target–student KL from 0.152 to 0.031. The resulting target reaches the endpoint arm’s best evaluated Math score with 35% fewer allocated H100-hours. Across our experiments, the results suggest that shift targets are particularly useful when teacher signals are combined at a state. With three composed teachers, Δ -MOPD exceeds endpoint composition by 4.11 Math and 1.95 five-benchmark points; with two, it matches endpoint accuracy. In an unpaired cross-tokenizer extension, a projected fourth shift improves all three Math benchmarks and adds a further 3.32 Math and 1.85 five-benchmark points. Under phased routing, it achieves higher mean performance in both phase orders and reduces the observed order gap from 10.50 to 6.42 points. Under interleaved routing, where each update involves one teacher, the two targets perform comparably. The phased results provide supporting evidence that the benefit may extend to signals accumulated across training phases. Target construction is thus an independent design axis in MOPD, complementary to teacher selection.

1 Introduction

On-policy distillation (OPD) trains a student on its own rollouts using scores from a frozen teacher, so supervision arrives at the states the student visits (Gu et al., 2024; Agarwal et al., 2024; Song and Zheng, 2026). Because post-trained specialists are individually incomplete and widely released, multi-teacher OPD (MOPD) aims to combine several of them in one student. MOPD is used in two settings. In *common-domain composition*, several teachers score the same rollout from one prompt domain and their signals are merged into one target. In *routed-domain distillation*, prompts from different domains are assigned to the corresponding specialist, so each state is supervised by one teacher.

Both settings, and the studies of routing, teacher counteraction, and cascaded overwrite built on them (Ma et al., 2026; Chen et al., 2026; Li et al., 2026), transfer each teacher’s *endpoint* policy. A specialist endpoint, however, combines the change produced by its post-training stage with preferences inherited from its base (DeepSeek-AI, 2025). When a teacher’s base differs from the student’s initialization, endpoint supervision transfers that inherited difference as well. The original MOPD pipeline avoids it because every teacher shares the student’s initial checkpoint (Ma et al., 2026); public specialists rarely do. Single-teacher methods report promising results from teacher-relative differences (Heo et al., 2026; Feng et al., 2026; Yang et al., 2026), but whether such differences are a better transfer object in either MOPD setting has not been tested.

We separate two decisions in any MOPD recipe: *teacher selection*, which teachers supervise a state, and *target construction*, how their scores define the distribution to be matched. Holding selection fixed, we compare endpoint targets with Δ -MOPD, which measures each teacher against its own base B_i and re-anchors the resulting shift at the student’s

initialization A . Writing z_X for checkpoint X 's logits at a student-generated state and $\mathcal{S}$ for the selected teachers, the two targets are

$$\begin{aligned} \text{endpoint: } z_E &= z_A + \sum_{i \in \mathcal{S}} (z_{T_i} - z_A), \\ \Delta\text{-MOPD: } z_C &= z_A + \sum_{i \in \mathcal{S}} (z_{T_i} - z_{B_i}), \end{aligned} \quad (1)$$

each normalized by a softmax. Composition selects all teachers; routing selects one. When $B_i = A$, both constructions give the same term, so only cross-origin teachers are affected (Figure 1). Concretely, if a Math specialist both acquires stronger reasoning and keeps a verbosity preference already present in its base, endpoint distillation transfers both; dividing the specialist by its base cancels what the base already preferred, and the anchor expresses the remaining change relative to the student's own starting policy. We hypothesize that removing inherited base differences yields a target that is closer to the student and, under composition, more balanced across teachers.

Our results suggest that shift targets are particularly useful when teacher signals are combined at a state. The mechanism appears directly in the learning signal: endpoint supervision gives the inherited base pull a larger gradient than the post-training shift, whereas base subtraction produces a target that is five times closer to the student and substantially more balanced across teachers. The closer target reaches the endpoint arm's best Math score with 35% fewer allocated H100-hours. In common-domain composition, shifts match endpoint accuracy with two teachers and lead on Math and the five-benchmark macro with three. In an unpaired cross-tokenizer extension, a projected fourth shift from a second cross-origin teacher improves all three Math benchmarks, raising the Math macro by a further 3.32 points and the five-benchmark macro by 1.85 points. In routed-domain distillation, phased routing yields higher mean performance for shifts in both phase orders and a smaller observed order gap; interleaved routing, where each update sees a single teacher, yields comparable results. The phased results provide supporting evidence that the benefit may extend to signals accumulated across training phases.

Contributions.

1. We identify target construction as a design axis in MOPD, independent of teacher selection, and instantiate it with Δ -MOPD (Eq. 1), which

reduces exactly to endpoint OPD for same-origin teachers.

2. **We expose the mechanism that impedes endpoint transfer.** The inherited base pull can exceed the post-training shift; removing it lowers the teacher-term norm ratio from 5.2:1 to 1.44:1, and reduces target-student KL fivefold.
3. **The improved geometry yields stronger signal combination.** With three composed teachers, Δ -MOPD exceeds endpoint composition by 4.11 Math and 1.95 five-benchmark points; under an unpaired cross-tokenizer extension, a projected fourth shift improves all three Math benchmarks and adds a further 3.32 Math and 1.85 five-benchmark points. Under phased routing it achieves higher mean performance in both orders and reduces the observed order gap from 10.50 to 6.42 points. Interleaved routing performs comparably.

2 Preliminaries

Let $x \sim \mathcal{D}$ and $y = (y_1, \dots, y_L) \sim \pi_\theta(\cdot | x)$, with $h_t = (x, y_{<t})$, $p_t = \pi_\theta(\cdot | h_t)$, and q_t the target distribution at h_t . OPD minimizes the token-level reverse KL

$$\mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot | x)} \left[\frac{1}{L} \sum_{t=1}^L D_{\text{KL}}(p_t \| q_t) \right]. \quad (2)$$

For sampled $y_t \sim p_t$, $\log p_t(y_t) - \log q_t(y_t)$ is a one-sample estimator of the per-state reverse KL. Training maximizes the sampled-token reward $\log q_t(y_t) - \log p_t(y_t)$ with the centered score-function estimator of Appendix B.

Notation. π_X and z_X denote checkpoint X 's distribution and logits over a shared effective vocabulary. A is a frozen anchor, in every experiment the student initialization; T_i is teacher i 's post-training endpoint; B_i is the exact checkpoint that entered that stage. The *shift* of teacher i is $\delta_i = z_{T_i} - z_{B_i}$ and its *inherited base difference* is $z_{B_i} - z_A$. A shift carries the complete effect of post-training—capability, style, and calibration alike—and is not assumed to isolate capability. Teacher i is *same-origin* when $B_i = A$ and *cross-origin* otherwise.

Teacher selection. A selection rule assigns each state a teacher set $\mathcal{S}(h_t)$. Common-domain composition uses $\mathcal{S}(h_t) = \{1, \dots, M\}$, so the teachers'

Each dot is a model. Thick arrows are the per-teacher signals; adding them tail-to-head gives the composite target (★) the student is trained toward.

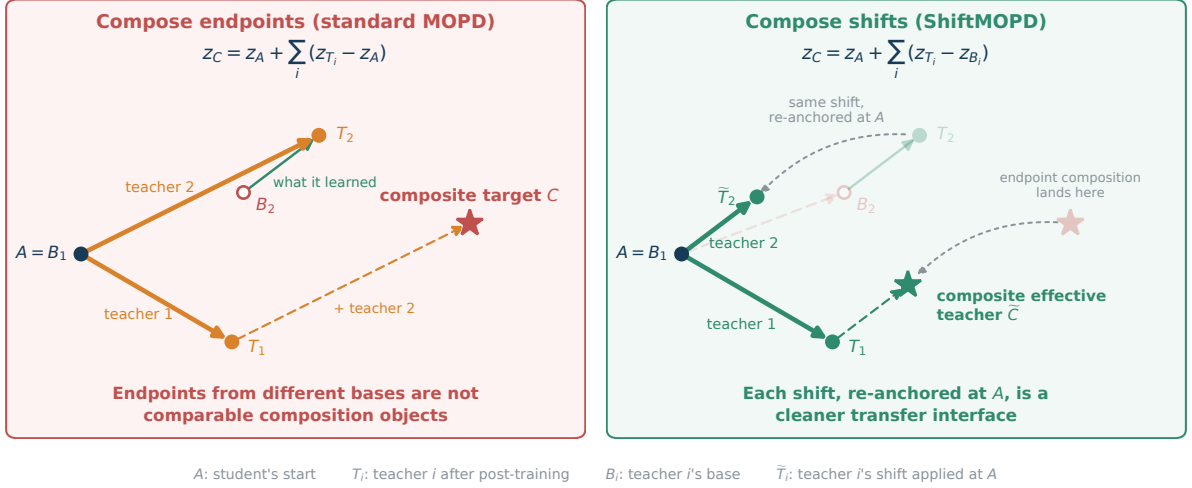


Figure 1: Endpoint versus shift targets. Each endpoint term $z_{T_i} - z_A$ contains a post-training shift $z_{T_i} - z_{B_i}$ and an inherited base difference $z_{B_i} - z_A$. Δ -MOPD keeps only the shift and re-anchors it at A , giving the effective teacher $\tilde{T}_i$; the two composite targets differ by the summed base differences. Teacher 1 is drawn with $B_1 = A$, so $\tilde{T}_1 = T_1$. Positions are schematic.

relative magnitudes at a state determine the update. Routed-domain distillation uses $\mathcal{S}(h_t) = \{\rho(h_t)\}$, where ρ maps a prompt to its domain’s teacher; domains may be interleaved within training or assigned to contiguous phases.

3 Δ -MOPD

3.1 Construction

For teachers that share a token-id mapping on the 151,665-token effective vocabulary, Δ -MOPD uses the shared-anchor target

$$\pi_C(v | h_t) = \frac{1}{Z(h_t)} \pi_A(v | h_t) \prod_{i \in \mathcal{S}(h_t)} \frac{\pi_{T_i}(v | h_t)}{\pi_{B_i}(v | h_t)}, \quad (3)$$

where $Z(h_t)$ normalizes over the vocabulary. Each model’s log-partition term is constant across tokens, so the target is a logit sum normalized once:

$$z_C(v | h_t) = z_A(v | h_t) + \sum_{i \in \mathcal{S}(h_t)} \delta_i(v | h_t), \quad (4)$$

$$\pi_C(\cdot | h_t) = \text{softmax}(z_C(\cdot | h_t)),$$

with sampled-token reward

$$R_t^C(v) = \log \pi_C(v | h_t) - \log \pi_\theta(v | h_t). \quad (5)$$

Setting $q_t = \pi_C(\cdot | h_t)$ in Eq. 2 gives the full-vocabulary objective. Some runs approximate only the partition function on a restricted support; details of this approximation are given in Appendix B.

Endpoint control. The control replaces each shift by $z_{T_i} - z_A$, giving z_E in Eq. 1. At a given prefix the two pre-softmax targets differ by exactly $\sum_{i \in \mathcal{S}(h_t)} (z_{B_i} - z_A)$. With one selected teacher the control is that teacher’s endpoint.

Same-origin reduction. If $B_i = A$, then $z_A + \delta_i = z_{T_i}$: a same-origin teacher’s shift target *is* its endpoint. Routing to a same-origin teacher is therefore exactly endpoint OPD, and if every selected teacher is same-origin the two composites coincide. Only cross-origin terms differ.

3.2 Reference-term interpretation

The endpoint reward of teacher j decomposes exactly as

$$\log \pi_{T_j} - \log \pi_\theta = \underbrace{\log \pi_{T_j} - \log \pi_{B_j}}_{R_j^\Delta: \text{shift}} + \underbrace{\log \pi_{B_j} - \log \pi_\theta}_{R_j^B: \text{base-reference pull}}. \quad (6)$$

The base-reference pull equals the inherited difference $\log \pi_{B_j} - \log \pi_A$ plus the current anchor-student gap $\log \pi_A - \log \pi_\theta$; at initialization only the inherited difference remains. Up to a token-independent normalizer, Δ -MOPD has the same

form with a single shared reference:

$$R_t^C(v) = \sum_{i \in \mathcal{S}(h_t)} \delta_i(v \mid h_t) + [\log \pi_A(v \mid h_t) - \log \pi_\theta(v \mid h_t)].$$

Both targets reward following teacher shifts and pull toward a reference. Endpoint supervision pulls toward each selected teacher’s own base; Δ -MOPD replaces these heterogeneous pulls by one pull toward the student’s initialization, which is zero at the start of training. Two consequences are plausible: the target is closer to the student, and, under composition, a cross-origin teacher’s inherited difference no longer sets the scale of its term relative to the others. Section 5.1 examines both.

4 Experimental Setup

Checkpoints. The anchor and student initialization in every run is DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025). Table 1 lists the teachers and the exact precursor used for each shift. Polaris-7B is post-trained from the 7B checkpoint of the same distillation family, so “cross-origin” here means a different and larger base of the same lineage. Nemotron and JustRL are same-origin, so their shift and endpoint terms are identical. In every paired comparison, the two targets therefore differ by one controlled substitution: the endpoint or teacher-relative representation of Polaris. A tokenizer-projected Qwen3-DAPO-449 shift extends the construction to a second cross-origin teacher in Appendix F.

Runs. *Common-domain composition* uses two runs on BigMath prompts, with every teacher scoring every rollout: the **mechanism** run composes Nemotron and Polaris ($M = 2$) and records diagnostics; the **scaling** run compares $M = 2$ with $M = 3$, which adds JustRL. *Routed-domain distillation* assigns Math prompts to Polaris and Science/IF prompts to Nemotron, either **interleaved** within each batch or in two **phased** orders. The **acquisition** run distills Polaris alone. Paired arms share initialization, prompts, optimizer, update budget, and evaluation; all frozen models score the same prefixes within an arm, although the arms’ on-policy trajectories diverge. Each comparison uses a shared run protocol; key configurations are listed in Table 4.

4.1 Diagnostic metrics

We measure $\Gamma_{\text{base}} = \|g_B\|/\|g_\Delta\|$, the gradient norm of the base-reference pull relative to the shift. For composition geometry, we report the teacher-term norm ratio, centered-shift cosine, cancellation, and top-16 sign conflict. Target–student and target–anchor KL are computed over the full effective vocabulary, independently of the training normalizer. The reported forward KL $D_{\text{KL}}(\pi_C \parallel \pi_\theta)$ measures target tracking and is not the training objective. Definitions are in Appendix B.

5 Results

The evidence follows a mechanism-to-outcome progression. We first test whether inherited base differences distort the learning signal (§5.1), then examine the predicted consequences for cross-origin acquisition (§5.2), common-domain composition (§5.3), and routed-domain distillation (§5.4).

5.1 Inherited base differences distort composition geometry

Inherited pull dominates early learning. Under endpoint supervision, the base-reference pull of Polaris has roughly twice the gradient norm of its shift ($\Gamma_{\text{base}} \approx 2$, averaged over checkpoints recorded every 20 steps). The ratio declines during training because only R_j^B depends on the student: it shrinks as the student moves toward π_{B_j} , whereas R_j^Δ is fixed by the teacher–base pair at a given state. The inherited component therefore has its greatest relative influence while the specialization is still being acquired.

Base subtraction restores balance. With the Nemotron term identical in both arms, the teacher-term norm ratio falls from 5.2 under endpoint composition to 1.44 under shift composition, and target–student KL falls from 0.152 to 0.031 (Figure 2). The shift composite lies near the equal-norm independence reference (cancellation 0.287 versus 0.293, cosine -0.016), whereas the endpoint composite has low cancellation (0.136), the signature of one dominant term. The ordering appears by step 20 and persists through training (Appendix C); at $M = 3$, target–anchor KL remains lower for shifts (0.483 versus 0.695).

Together, these measurements establish the proposed mechanism: inherited base differences enlarge the target and set the relative scale of the teacher terms. Removing them yields two testable predictions. The closer target should be acquired

Table 1: Public checkpoints. “Base” is the exact precursor used to compute each shift. Nemotron is pinned to revision v1.

Role	Model	Base	Origin	Specialization	Use
Anchor / student init	DS-R1-Distill-Qwen-1.5B	self	self	–	all runs
Math teacher	JustRL-DeepSeek-1.5B	DS-R1-Distill-Qwen-1.5B	same	Math RL	composition ($M = 3$)
Science/IF teacher	Nemotron-Reasoning-Qwen-1.5B	DS-R1-Distill-Qwen-1.5B	same	multi-domain RL	composition, routing
Math teacher	Polaris-7B-Preview	DS-R1-Distill-Qwen-7B	cross	Math RL	all runs
Math teacher	Qwen3-DAPO-449	Qwen3-8B	cross	DAPO Math RL	extension (App. F)

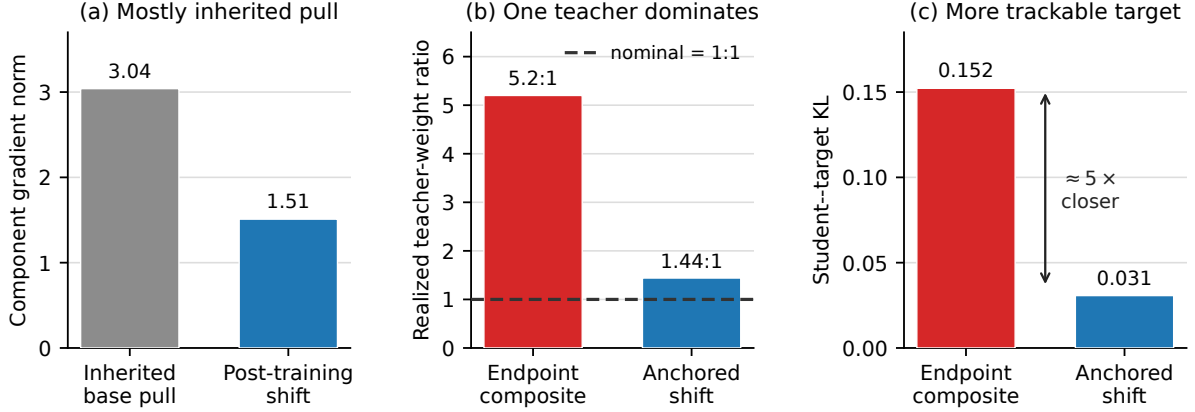


Figure 2: Composition diagnostics in the mechanism run. **(a)** For Polaris, the base-reference pull has a larger gradient norm than the shift. **(b)** The endpoint composite has a 5.2:1 teacher-term norm ratio; the shift composite has 1.44:1. **(c)** The shift target is about five times closer to the student in target–student KL. Values are late-training averages (Table 7).

earlier, and the more balanced terms should make additional teachers more useful. We test these predictions next.

5.2 A closer target accelerates cross-origin acquisition

Distilling Polaris alone removes composition and directly tests the first prediction. Δ -MOPD leads the English-Math macro by 2.87 pp at step 100, after which the arms approach a similar level. Including precursor-scoring overhead, Δ -MOPD scores 47.90% at step 100, already exceeding Endpoint’s best evaluated 46.54% at step 195; after charging for Δ -MOPD’s slower updates, these checkpoints cost 28.0 versus 42.9 allocated H100-hours. At initialization, the endpoint target is $4.94\times$ farther from the student in target–student KL across 16,384 shared prefix states. The closer target is therefore acquired earlier even after charging for the additional precursor forward pass. Details are in Appendix E.

5.3 Balanced shifts improve common-domain composition

We next test the second prediction: whether balancing the teacher terms turns additional supervision

into usable capability.

Two teachers. In the mechanism run, both composites improve substantially over the initial student and reach similar aggregate accuracy: 44.91% for Δ -MOPD versus 45.14% for the endpoint composite under Avg@K, and 37.73% versus 36.25% under greedy decoding. Δ -MOPD is higher on Science/IF under both decodings, while the endpoint composite is 3.4 pp higher on MATH-500 (Appendix C).

Three teachers. The scaling run compares $M = 2$ with $M = 3$ under one protocol (Table 2). At $M = 3$, Δ -MOPD exceeds the endpoint composite by 4.11 pp on the Math macro and 1.95 pp on the five-suite macro. Adding the same JustRL teacher raises shift composition by 4.40 Math and 1.78 five-suite pp, but endpoint composition by only 0.43 and 0.58. Because JustRL is same-origin, its term is identical in both arms; the difference lies in how Polaris is represented when a third teacher is added. The gain is concentrated in AIME 2025; Science/IF, which receives no in-domain prompts in this run, is 1.29 pp below the endpoint composite (Table 9).

Shift composition thus preserves two-teacher ac-

Table 2: Common-domain composition in the scaling run at step 100. Greedy pass@1 macros (%). All teachers score every BigMath rollout. Δ_{3-2} is the change from adding same-origin JustRL. Per-benchmark results are in Table 9.

Teachers	Target	Math	Sci/IF	Five-suite
$M = 2$	Endpoint composite	28.96	15.70	23.66
$M = 2$	Δ -MOPD	29.10	17.38	24.41
$M = 3$	Endpoint composite	29.39	16.51	24.24
$M = 3$	Δ -MOPD	33.50	15.22	26.19
Δ_{3-2}	Endpoint composite	+0.43	+0.81	+0.58
Δ_{3-2}	Δ -MOPD	+4.40	-2.16	+1.78

Table 3: Routed-domain distillation. Interleaved routing mixes domains in each batch (50/25/25 Math/Science/IF); phased routing trains one teacher-domain phase at a time. Greedy pass@1 macros (%); Δ is Δ -MOPD minus Endpoint in pp.

Schedule	Macro	Endpoint	Δ -MOPD	Δ
Interleaved	Math	30.61	31.52	+0.91
	Sci/IF	18.26	18.45	+0.19
	Five-suite	25.67	26.29	+0.62
Phased: Sci/IF $\rightarrow$ Math	Math	28.02	34.39	+6.37
	Sci/IF	18.25	22.89	+4.64
	Five-suite	24.11	29.79	+5.68
Phased: Math $\rightarrow$ Sci/IF	Math	40.87	42.20	+1.33
	Sci/IF	25.23	27.23	+2.00
	Five-suite	34.61	36.21	+1.60

curacy and turns an additional teacher into a substantially larger Math gain.

5.4 Shift targets show consistent gains under phased routing

Composition combines teacher signals within each state. Routing separates them across examples or training phases, examining whether the same target construction also changes how specialist signals accumulate over time. Under routing, Polaris supervises Math prompts and Nemotron supervises Science/IF prompts. Because Nemotron is same-origin, the two targets coincide on Science/IF states and differ only in the routed Polaris term. Table 3 reports two schedules.

Interleaved routing. With both teachers active throughout training, the five-suite macro differs by only 0.62 pp in favor of Δ -MOPD, with group-macro differences of +0.91 pp on Math and +0.19 pp on Science/IF. The two targets are therefore comparable when each update involves a single teacher, as expected if the benefit of shifts comes from reconciling several teacher signals.

Phased routing. When each teacher-domain pair occupies a contiguous phase, Δ -MOPD ends above its same-order endpoint control in both orders, by 5.68 and 1.60 five-suite pp, and both group macros improve in each order. The gap between the two phase orders falls from 10.50 pp under endpoint supervision to 6.42 pp. Phased training couples each teacher with its domain data, providing evidence about the robustness of the combined teacher-data curriculum across phase orders.

Together, these results suggest that the benefits of shift targets may extend beyond within-state composition to settings where teacher signals accumulate across training phases. We view the phased-routing results as supporting evidence for this interpretation, while interleaved routing yields comparable performance.

6 Discussion

Where shifts help. Shift targets show their clearest benefits when several teachers are composed at a state. Phased routing provides supporting evidence that these benefits may extend to settings where routed teachers occupy separate phases and their signals accumulate over training; under interleaved routing, where each update involves one teacher, the targets are comparable. The paired design holds all same-origin terms fixed and changes only the representation of Polaris, directly attributing the contrast to target construction. The projected four-teacher extension adds a second cross-origin shift, improves all three Math benchmarks, and raises the Math macro by a further 3.32 pp (Appendix F), extending the construction to a larger heterogeneous teacher set.

Geometry and scale. Cosine and sign conflict are invariant to rescaling individual shifts, and cancellation to rescaling all shifts jointly. Their change therefore reflects the relative geometry of the teacher terms, not only the overall distance to the student. The policy-ratio form cancels vocabulary normalization constants but does not calibrate differences in logit temperature or confidence across checkpoints, so shift norms reflect both the magnitude of post-training change and the scale on which it is expressed. Calibration-aware composition is complementary to base subtraction.

7 Related Work

Distillation for language models. Knowledge distillation transfers a teacher’s output distribution

to a smaller student (Hinton et al., 2015), for language models mostly off-policy on teacher-written rationales (Li et al., 2022; Hsieh et al., 2023; Fu et al., 2023; Singh et al., 2024; Liu et al., 2024), so the student is never supervised at the states it visits. OPD removes that mismatch by scoring the student’s own rollouts (Agarwal et al., 2024; Song and Zheng, 2026) and has been reframed as compressing reward-shaped behavior (Sang et al., 2026; Xu et al., 2026; He et al., 2026; Xu et al., 2026). Work that broadens what the on-policy signal carries, through rubric judgments (Fang et al., 2026), dual sequence- and token-level views (Hou et al., 2026; Wang et al., 2026; Shen et al., 2026), or context rather than weights (Ye et al., 2026), keeps the teacher’s endpoint as the matched object. Li et al. (2026) show that single-teacher OPD depends on thinking-pattern compatibility rather than teacher strength; student competence frontiers for on-policy self-distillation are studied by Xu et al. (2026).

Multi-teacher on-policy distillation. MOPD studies of routing, counteraction, and cascaded overwrite (Ma et al., 2026; Chen et al., 2026; Li et al., 2026) motivate our paired controls, and the same brittleness surfaces in tool-use, GUI-agent, and expert-to-generalist settings (Shen et al., 2026; Lian et al., 2026; Chen et al., 2026). Sample routing connects group-relative and self-distillation policy optimization (Li et al., 2026). These methods act on selection, scheduling, or rehearsal; we vary what each selected teacher contributes.

Differences as transfer objects. Delta-oriented single-teacher methods use teacher–base differences for direct or extrapolated transfer (Heo et al., 2026; Feng et al., 2026; Yang et al., 2026), and task arithmetic and proxy tuning compose parameter- or logit-space differences offline (Ilharco et al., 2023; Liu et al., 2024). We study such differences as transfer objects in both MOPD settings, on the student’s own prefixes and normalized under one softmax.

8 Conclusion

We studied target construction in MOPD by comparing teacher-relative shifts, re-anchored at the student’s initialization, with endpoint supervision. Inherited base differences can dominate a teacher’s post-training shift; removing them yields a target that is balanced across teachers and closer to

the student. Across our experiments, shift targets show their clearest benefits when teacher signals are combined at a state: they match endpoints with two composed teachers and raise Math and five-benchmark accuracy with three. In an unpaired cross-tokenizer extension, adding a projected fourth shift from a second cross-origin teacher yields further gains on both macros. Under phased routing, shifts achieve higher mean performance in both tested orders and reduce the observed order gap, providing supporting evidence that their benefits may extend to teacher signals accumulated across phases. Under interleaved routing, where each update sees one teacher, the targets are comparable. Because the construction reduces exactly to endpoint OPD for same-origin teachers, it preserves their signal and acts only on inherited base differences. What is transferred from each teacher is thus a design axis in MOPD, complementary to which teachers are selected.

9 Limitations

- **Evaluation scope.** We study public reasoning specialists and common-domain composition on Math prompts, with Science/IF measured through held-out evaluation. Broader specialist families and domain-mixed composition are natural extensions.
- **Requirements.** Δ -MOPD requires each teacher’s precursor and a compatible tokenizer or an explicit projection rule, unlike black-box OPD (Ye et al., 2025).

References

- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*.
- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*.
- Mingyang Song and Mao Zheng. 2026. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*.
- Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. 2026. **MOPD: Multi-teacher on-policy**

- distillation for capability integration in LLM post-training. *Preprint*, arXiv:2606.30406.
- Tianlei Chen, Jiao Ou, Ziyuan Liu, Ruiming Tang, Jian Liang, and Han Li. 2026. [Counteraction-aware multi-teacher on-policy distillation for general capability recovery with domain preservation](#). *Preprint*, arXiv:2605.27115.
- Zhaoyi Li, Deyang Kong, Yuan Wei, Evan Yang, Ranran Shen, Mahardika Krisna Ihsani, Ming Yang, Wei Zhang, Chuan Hao, Jian Yang, Ran Tao, Bryan Dai, Shikun Zhang, Wei Ye, Ying Wei, and Defu Lian. 2026. [Every coin has two sides: On the dual nature of generalization in on-policy distillation of large language models](#). *Preprint*, arXiv:2608.16647.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638.
- Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. 2026. [On-policy delta distillation](#). *Preprint*, arXiv:2607.15161.
- Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. 2026. [Weak-to-strong generalization via direct on-policy distillation](#). *Preprint*, arXiv:2607.05394.
- Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026. [Learning beyond teacher: Generalized on-policy distillation with reward extrapolation](#). *Preprint*, arXiv:2602.12125.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhui Chen, and Xin Xie. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*.
- Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.
- Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, J. Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, and 22 others. 2024. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*.
- Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Zhiqi Bai, Jie Liu, Ge Zhang, Jiakai Wang, Yanan Wu, Congnan Liu, Wenbo Su, Jiamang Wang, Lin Qu, and Bo Zheng. 2024. DDK: Distilling domain knowledge for efficient large language models. In *Advances in Neural Information Processing Systems*.
- Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. 2026. [CRISP: Compressed reasoning via iterative self-policy distillation](#). *Preprint*, arXiv:2603.05433.
- Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. 2026. [TIP: Token importance in on-policy distillation](#). *Preprint*, arXiv:2604.14084.
- Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. 2026. [Self-distillation zero: Self-revision turns binary rewards into dense supervision](#). *Preprint*, arXiv:2604.12002.
- Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. 2026. [Beyond GRPO and on-policy distillation: An empirical sparse-to-dense reward principle for language-model post-training](#). *Preprint*, arXiv:2605.12483.
- Junfeng Fang, Zhepei Hong, Mao Zheng, Mingyang Song, Gengsheng Li, Houcheng Jiang, Dan Zhang, Haiyun Guo, Xiang Wang, and Tat-Seng Chua. 2026. Rubric-based on-policy distillation. *arXiv preprint arXiv:2605.07396*.
- Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. 2026. Uni-OPD: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*.
- Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. 2026. [Not all disagreement is learnable: Token teachability in on-policy distillation](#). *Preprint*, arXiv:2605.26844.
- Guobin Shen, Lei Huang, Xiang Cheng, Chenxiao Zhao, Jindong Li, Dongcheng Zhao, and Xing Yu. 2026. [From generic correlation to input-specific credit in on-policy self-distillation](#). *Preprint*, arXiv:2605.11613.
- Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2026. [On-policy context distillation for language models](#). *Preprint*, arXiv:2602.12275.

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

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. 2026. [PACED: Distillation and on-policy self-distillation at the frontier of student competence](#). *Preprint*, arXiv:2603.11178.

Jiabin Shen, Guang Chen, and Chengjun Mao. 2026. [Diagnosing and calibrating tool-call boundary drift in multi-teacher on-policy distillation](#). *Preprint*, arXiv:2607.07050.

Niu Lian, Alan Chen, Zhehao Yu, Chengzhen Duan, Fazhan Liu, Hui Liu, Pei Fu, Jian Luan, Yaowei Wang, Shu-Tao Xia, and Jinpeng Wang. 2026. [UI-MOPD: Multi-platform on-policy distillation for continual GUI agent learning](#). *Preprint*, arXiv:2607.04425.

Yunjie Chen, Xiaoxin Chen, and Fang Wang. 2026. [RE-GEN: Replay-recycling for expert-to-generalist distillation with offline reinforcement learning](#). *Preprint*, arXiv:2607.19450.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. 2026. [Unifying group-relative and self-distillation policy optimization via sample routing](#). *Preprint*, arXiv:2604.02288.

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

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. Tuning language models by proxy. In *International Conference on Learning Representations*.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shao-han Huang, and Furu Wei. 2025. [Black-box on-policy distillation of language models](#). *Preprint*, arXiv:2511.10643.

A Protocols and Reproducibility

Evaluation. Evaluation uses SGLang with a 16,384-token maximum response length, except that IF-Eval is capped at 2,048 tokens by the evaluation driver; the scaling run and teacher references use a 10,000-token greedy protocol. Greedy decoding uses temperature 0 and top- $p = 1$; Avg@ K sampling uses temperature 1 and top- $p = 0.95$. MATH-500 reuses its greedy item outcomes for $K = 1$. All runs use the same 1,309 frozen held-out items. Five-suite macros are unweighted averages of the five benchmark scores, not of the two

group macros. Entries with error bars are mean $\pm$ sample standard deviation over five independently trained seeds with matched seed IDs and prompt orders; macros are computed from the five-seed benchmark means.

Reproducibility. Each teacher is paired with its exact precursor and checked for tokenizer compatibility; the Qwen3 extension records the token-string overlap map of Eq. 16. Paired arms share prompts, update counts, and evaluation settings, and all frozen models score identical student prefixes within an arm before target construction. Dataset revisions, sampled IDs, chat templates, model revisions, decoding settings, and verifier versions are fixed across comparisons.

B Estimator and Diagnostic Definitions

Estimator and normalization. For valid sampled-token positions $\mathcal{I}$ in an optimization batch, training subtracts the empirical mean reward $\bar{R}^C = |\mathcal{I}|^{-1} \sum_{(b,t) \in \mathcal{I}} R_{b,t}^C(y_{b,t})$ and optimizes the detached centered reward times the student log-probability. This is sampled-token centering rather than an explicit per-state full-vocabulary expectation. When a restricted support S_t is used, the sampled-token numerator is exact and only the composite partition is approximated:

$$\begin{aligned} \widehat{\log \pi_C}(y_t | h_t) &= z_C(y_t | h_t) \\ &\quad - \log \sum_{u \in S_t} \exp z_C(u | h_t). \end{aligned} \quad (7)$$

If $m_{S_t} = \sum_{v \in S_t} \pi_C(v | h_t)$ is the composite mass captured by the support, then $\widehat{\log \pi_C}(y_t | h_t) = \log \pi_C(y_t | h_t) - \log m_{S_t}$, so the restricted normalizer raises the sampled-token log probability by $-\log m_{S_t}$. This is a normalizer approximation, not the top- k OPD objective. All logits are restricted to the 151,665-token effective vocabulary before composition.

Gradient decomposition. For sampled positions $\{(h_t, y_t)\}_{t=1}^N$, define $R_t^A = \log \pi_T(y_t | h_t) - \log \pi_B(y_t | h_t)$ and $R_t^B = \log \pi_B(y_t | h_t) - \log \pi_\theta(y_t | h_t)$. Their gradient estimators and rela-

Table 4: Training protocols. Values are shared by both arms of a run unless an arm-specific value is shown. Fields marked “–” were not recorded in the run’s frozen configuration summary.

Setting	Acquisition ($M = 1$)	Mechanism ($M = 2$)	Scaling ($M = 2, 3$; proj. $M = 4$)	Phased routing	Interleaved routing
Phase structure	single	single	single	2 phases, optimizer resumed, dataloader reset	single, domains interleaved from step 1
On-policy prompts	BigMath, 25K Math	BigMath, 25K Math	BigMath Math	BigMath; TextbookReasoning + Llama-Nemotron IF	Math + Science/IF mixture (50/25/25)
Temperature / top- p	1.0 / 1.0	1.0 / 1.0	1.0 / 1.0	1.0 / 1.0	1.0 / 1.0, top- k off
Maximum response	10,000	10,000	10,000	10,000	10,000
Learning rate	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
Hardware	$4 \times$ H100 80GB	$4 \times$ H100 80GB	$8 \times$ H200	$4 \times$ H100 80GB per arm	$8 \times$ H200 (Endpoint) / $8 \times$ H100 (Δ -MOPD)
Evaluation	Avg@ K (Table 5)	Avg@ K and greedy	greedy pass@1	greedy pass@1	greedy pass@1

Table 5: Held-out evaluation suite. Avg@ K applies to the acquisition and mechanism runs; other runs decode every benchmark greedily at $K = 1$.

Group	Benchmark	Test items	Avg@ K	Generation / scoring
English Math	AMC 2023	40	16	sampled, official answer checker
English Math	MATH-500	500	1	greedy, official answer checker
English Math	AIME 2025	30	16	sampled, official answer checker
Science	GPQA-Diamond	198	4	sampled, final option-letter scorer
Instruction following	IF-Eval	541	4	sampled, vendored rule checker

tive magnitude are

$$g_{\bullet} = \nabla_{\theta} \left[-\frac{1}{N} \sum_{t=1}^N \text{sg}(R_t^{\bullet}) \log \pi_{\theta}(y_t | h_t) \right],$$

$$\bullet \in \{\Delta, B\},$$

$$\Gamma_{\text{base}} = \frac{\|g_B\|_2}{\|g_{\Delta}\|_2 + 10^{-8}}. \quad (8)$$

At each mechanism checkpoint, Eq. 8 uses the cross-origin teacher–base pair, the first micro-batch, and its first $N = \min(128, |\mathcal{I}|)$ valid response positions, backpropagating through all trainable parameters. Measurements are recorded every 20 steps from step 20 through 380.

Composition geometry. Center each shift by its vocabulary mean,

$$\Delta_i(t, v) = z_{T_i}(v | h_t) - z_{B_i}(v | h_t)$$

$$- \frac{1}{V} \sum_{u=1}^V [z_{T_i}(u | h_t) - z_{B_i}(u | h_t)], \quad (9)$$

concatenate across sampled positions, and compute pairwise cosine, aggregate norm, and

$$\text{Cancel}(\mathcal{T}) = 1 - \frac{\|\sum_{i \in \mathcal{T}} \Delta_i\|_2}{\sum_{i \in \mathcal{T}} \|\Delta_i\|_2 + 10^{-8}}. \quad (10)$$

Sign conflict uses the top-16 absolute-shift coordinates selected by each teacher, retaining overlaps and aggregating across positions. The endpoint control replaces $z_{T_i} - z_{B_i}$ by $z_{T_i} - z_A$. Because Δ_i is evaluated at student-visited prefixes, these statistics depend on the rollout protocol and pooling rule. The mechanism run concatenates the first 128 valid positions of the first micro-batch on each data-parallel rank; the scaling run computes the same quantities per complete response and averages over micro-batches. We read them within a run.

Two orthogonal equal-norm vectors have cosine 0 and $\text{Cancel} = 1 - \sqrt{2}/2 = 0.2929$; independent sign-symmetric coordinates have conflict rate 0.5. For $a = \Delta_{\text{same}}$, $b = \Delta_{\text{cross}}$, $c = \cos(a, b)$, and

$$r = \|b\|/\|a\|,$$

$$1 - \text{Cancel} = \frac{\sqrt{r^2 + 1 + 2rc}}{r + 1}. \quad (11)$$

Since this expression is invariant under $r \mapsto 1/r$, cosine and cancellation identify the norm ratio $\max(r, 1/r)$.

Target distance. The target-to-student forward KL is

$$D_{\text{ST}} = \mathbb{E}_{h_t} \sum_{v=1}^V \pi_C(v \mid h_t) \log \frac{\pi_C(v \mid h_t)}{\pi_\theta(v \mid h_t)}, \quad (12)$$

computed over the full $V = 151,665$ vocabulary on the first micro-batch of each mechanism checkpoint, using up to 128 valid response positions. KL to the anchor replaces π_θ by π_A .

C Common-Domain Composition Details

C.1 Mechanism run: capability

Under greedy decoding, the group-balanced totals are 24.10% for the initial student, 37.73% for Δ -MOPD, and 36.25% for the endpoint composite. The per-group greedy macros are 31.44%, 50.91%, and 48.16% for English-Math, and 16.76%, 24.54%, and 24.34% for Science/IF, so greedy decoding favors Δ -MOPD on both groups while the Avg@ K ordering is effectively tied.

C.2 Mechanism run: diagnostics

Table 7 reports the late-training window and Table 8 early checkpoints; the two arms separate by step 20 and the ordering is stable.

C.3 Scaling run

Under the same greedy, 10,000-token protocol, the individual teachers score (Math/Sci-IF/five-suite macro, %) 46.62/28.08/39.21 for Nemotron-1.5B, 49.02/27.50/40.41 for JustRL-1.5B, and 59.29/36.87/50.32 for Polaris-7B.

At $M = 2$ the two shifts have norms 13,674 and 6,073 with pairwise cosine 0.084. Adding JustRL gives norms 12,891, 11,114, and 6,490, so the third shift enters at a magnitude comparable to the other two. Two of its pairwise cosines remain near zero (0.072 and 0.062) while the third rises to 0.315. Across the recorded $M = 2$ and $M = 3$ training-signal diagnostics, shift composition lowers target displacement, reward variance,

and gradient magnitude; at $M = 3$, its final target-to-anchor KL is 0.483 rather than 0.695. Its per-update time is 11.0% higher (167.1 versus 150.6 seconds) because the Polaris precursor must also be scored.

D Routed-Domain Distillation Details

Interleaved routing. The router is the prompt’s domain: Math prompts are supervised by Polaris and Science and instruction-following prompts by Nemotron, with exactly one teacher per prompt. The training mixture is 50% Math, 25% Science, and 25% instruction-following. Domains are interleaved from the first update with a deterministic within-batch shuffle. Endpoint matches $z_{T_{\rho(h)}}$ and Δ -MOPD matches $z_A + \delta_{\rho(h)}$. Targets are normalized exactly over the full effective vocabulary. The Endpoint arm was trained on $8 \times \text{H200}$ and the Δ -MOPD arm on $8 \times \text{H100}$; the comparison is at matched update count, so hardware affects throughput but not the reported accuracies. All checkpoints were evaluated on H100 with greedy decoding.

D.1 Phased routing

Each arm trains two phases with actor and optimizer state carried across the switch. The Science/IF phase is identical across methods, so the contrast lies in the Polaris Math phase. Per-item binary outcomes are retained, so same-order Δ -MOPD-versus-Endpoint comparisons are exact paired McNemar tests on identical items. The only comparison below $p = 0.05$ is MATH-500 in Math $\rightarrow$ Science/IF (56 versus 36 discordant items, $p = 0.047$), which does not survive correction across the ten benchmark-by-order comparisons.

E Cross-Origin Acquisition and Compute

Both arms train only on BigMath prompts. The English-Math macro across the evaluated checkpoints is 47.90/45.65/47.68 for Δ -MOPD and 45.03/46.18/46.54 for Endpoint at steps 100/150/195. On GPQA-Diamond and IF-Eval, the ordering flips with decoding mode: at step 100, Δ -MOPD leads by 0.91 pp on the Avg@4 macro while greedy decoding favors Endpoint by 1.07 pp. The endpoint target remains near $5 \times$ farther from the student in target-student KL across checkpoints.

Table 6: Mechanism run ($M = 2$) student evaluation. AMC 2023 and AIME 2025 use Avg@16, GPQA-Diamond and IF-Eval use Avg@4, and MATH-500 uses greedy decoding. Entries with error bars are mean $\pm$ sample standard deviation over training seeds, in percent.

Benchmark	Initial student	Δ -MOPD	Endpoint composite	Δ -MOPD – Endpoint (pp)
AMC 2023	63.12	79.84 ± 1.99	79.53 ± 2.79	+0.31
MATH-500	36.00	59.40 ± 1.29	62.80 ± 1.34	−3.40
AIME 2025	23.33	32.92 ± 3.04	32.92 ± 3.24	0.00
English Math macro	40.82	57.39	58.42	−1.03
GPQA-Diamond	34.72	40.28 ± 2.11	39.02 ± 2.22	+1.26
IF-Eval	23.43	24.58 ± 1.31	24.68 ± 1.07	−0.10
Science/IF macro	29.08	32.43	31.85	+0.58

Table 7: Late-training diagnostics for the mechanism run, averaged over steps 300, 320, 340, 360, 380.

Target	Cancellation	Aggregate norm	Conflict rate	Signal cosine	Entropy	KL to anchor	Target–student KL
Δ -MOPD	0.2871	3247.5	0.5540	−0.0163	0.2458	0.1243	0.0307
Endpoint composite	0.1363	10231.5	0.4386	0.0592	0.2643	0.2567	0.1522
<i>Equal-norm reference</i>	0.2929	—	0.5000	0.0000	—	—	—

Table 8: Early mechanism-run diagnostics at steps 20, 60, 100.

Target	Step	Cancellation	Target–student KL
Δ -MOPD	20	0.2728	0.1070
Δ -MOPD	60	0.2914	0.0750
Δ -MOPD	100	0.2832	0.0462
Endpoint composite	20	0.1265	0.2472
Endpoint composite	60	0.1378	0.1222
Endpoint composite	100	0.1381	0.1632
<i>Reference</i>	—	0.2929	—

E.1 Compute accounting

For an arm with S updates, measured end-to-end step times t_s , and N_{GPU} H100 GPUs, allocated compute is

$$C_{\text{alloc}} = \frac{N_{\text{GPU}}}{3600} \sum_{s=1}^S t_s \quad \text{H100-hours.} \quad (13)$$

Thus, one wall-clock hour of training on four allocated H100s counts as four H100-hours. The timed region includes rollout generation, student optimization, batching, and every teacher, precursor, and anchor forward pass, but excludes held-out evaluation. This measures allocated device time rather than FLOPs and does not correct for utilization. The $M = 1$ run uses each arm’s mean step time and four GPUs; the $M = 2$ run sums its recorded step times before multiplying by four GPUs.

For the acquisition comparison, Δ -MOPD first exceeds Endpoint’s best evaluated English-Math macro at step 100, whereas Endpoint attains that best score at step 195. With mean step times of

approximately 252 and 198 seconds, respectively, their allocated costs are

$$\begin{aligned} C_{\Delta\text{-MOPD}} &= 4 \times 100 \times 252/3600 = 28.0, \\ C_{\text{Endpoint}} &= 4 \times 195 \times 198/3600 = 42.9 \end{aligned} \quad \text{H100-hours.} \quad (14)$$

Therefore the matched-capability reduction is $(42.9 - 28.0)/42.9 = 34.7\%$, reported as 35%. The reduction comes from requiring fewer updates, not from cheaper updates: Δ -MOPD is 27.3% slower per update because it also scores the precursor.

At equal update count, Δ -MOPD costs 27.3% more in $M = 1$ and 12.0% more in $M = 2$, below a naive factor of two because rollout generation, optimization, batching, and parallel scoring are shared. Every H100-hour figure, including Table 12, already charges Δ -MOPD for this overhead. In the separate scaling run, the $M = 3$ overhead is 11.0% on the same eight-H200 configuration in both arms; this is a within-run wall-clock ratio, not an H200-to-H100 conversion.

Convergence criterion. We define the plateau as the first 25-step block after which every later block-mean rollout accuracy varies by at most 2.5 pp. Δ -MOPD stabilizes near step 75 and Endpoint near step 125, corresponding to 21.0 versus 27.5 H100-hours ($1.31\times$). At 21.0 hours Δ -MOPD has 53.38% block-mean rollout accuracy, against 51.12% for Endpoint at a comparable 22.0 hours. Best-macro compute, updates to plateau, and accuracy at matched hours therefore agree.

Table 9: Scaling run at step 100. Greedy pass@1; entries with error bars are mean $\pm$ sample standard deviation over training seeds, in percent.

Teachers	Target	AMC 2023	MATH-500	AIME 2025	Math macro	GPQA-D	IF-Eval	Sci/IF macro	Five-suite macro
$M = 2$	Endpoint composite	35.50 $\pm$ 2.56	34.04 $\pm$ 1.37	17.33 $\pm$ 1.93	28.96	16.77 $\pm$ 2.27	14.64 $\pm$ 1.66	15.70	23.66
$M = 2$	Δ -MOPD	38.00 $\pm$ 1.68	28.64 $\pm$ 1.42	20.67 $\pm$ 3.11	29.10	20.30 $\pm$ 1.90	14.45 $\pm$ 1.33	17.38	24.41
$M = 3$	Endpoint composite	38.00 $\pm$ 1.80	32.84 $\pm$ 0.99	17.33 $\pm$ 3.08	29.39	17.27 $\pm$ 1.53	15.75 $\pm$ 1.14	16.51	24.24
$M = 3$	Δ -MOPD	38.00 $\pm$ 1.61	31.84 $\pm$ 0.99	30.67 $\pm$ 2.50	33.50	17.27 $\pm$ 1.48	13.16 $\pm$ 1.50	15.22	26.19

Table 10: Phased routing at the step-300 checkpoint. Greedy pass@1; group macros are unweighted averages within each group. Entries with error bars are mean $\pm$ sample standard deviation over training seeds, in percent.

Order	Method	AMC 2023	MATH-500	AIME 2025	GPQA-Diamond	IF-Eval	Math macro	Sci/IF macro
Science/IF $\rightarrow$ Math	Endpoint	25.50 $\pm$ 2.43	41.24 $\pm$ 0.84	17.33 $\pm$ 2.19	15.76 $\pm$ 1.93	20.74 $\pm$ 0.81	28.02	18.25
	Δ -MOPD	33.00 $\pm$ 2.06	42.84 $\pm$ 1.31	27.33 $\pm$ 2.69	24.85 $\pm$ 2.28	20.92 $\pm$ 1.92	34.39	22.89
Math $\rightarrow$ Science/IF	Endpoint	40.50 $\pm$ 1.79	51.44 $\pm$ 0.94	30.67 $\pm$ 2.21	29.90 $\pm$ 1.81	20.55 $\pm$ 1.35	40.87	25.23
	Δ -MOPD	40.50 $\pm$ 2.48	55.44 $\pm$ 1.12	30.67 $\pm$ 2.52	32.42 $\pm$ 2.43	22.03 $\pm$ 1.47	42.20	27.23

Table 11: Acquisition run at step 100. AMC 2023 and AIME 2025 use Avg@16; MATH-500 uses greedy decoding. The macro is the unweighted average of these three benchmarks. Entries with error bars are mean $\pm$ sample standard deviation over training seeds, in percent.

Target	AMC 2023	MATH-500	AIME 2025	English Math macro
Δ -MOPD	67.81 $\pm$ 1.53	48.80 $\pm$ 1.40	27.08 $\pm$ 1.82	47.90
Endpoint	64.84 $\pm$ 1.71	44.20 $\pm$ 0.93	26.04 $\pm$ 1.73	45.03

Table 12: Earliest evaluated checkpoint meeting each held-out English-Math threshold in the acquisition run, in allocated H100-hours. Checkpoints were evaluated every 50 steps, so every entry is an upper bound on first-passage compute.

English-Math macro	Δ -MOPD	Endpoint
46.18%	$\leq$ 28.0 h	$\leq$ 33.0 h
46.54%	$\leq$ 28.0 h	$\leq$ 42.9 h
47.90%	$\leq$ 28.0 h	not observed by 42.9 h

F Cross-Tokenizer Extension

This extension adds Qwen3-DAPO-449, our public Qwen3-8B-DAPO iter-449 checkpoint, as a second cross-origin teacher in the scaling run ($M = 4$). It is tokenizer-identical to its Qwen3-8B precursor but not fully special-token compatible with the DeepSeek/Qwen2 anchor.

Overlap projection. We compute the Qwen3 shift inside the Qwen3 family and transfer only overlapping action coordinates to the anchor vocabulary. Let V_A be the anchor vocabulary, V_Q the Qwen3 vocabulary, and τ_A, τ_Q their id-to-token-string maps. With

$$\begin{aligned} \delta_Q^Q(v_Q | h_t) &= z_{\text{DAPO}}^Q(v_Q | h_t) \\ &\quad - z_{\text{Qwen3-8B}}^Q(v_Q | h_t), \end{aligned} \quad (15)$$

define $m(v_A) = v_Q$ whenever $\tau_A(v_A) = \tau_Q(v_Q)$. The projected shift is

$$\tilde{\delta}_Q(v_A | h_t) = \begin{cases} \delta_Q^Q(m(v_A) | h_t), & m(v_A) \in V_Q, \\ 0, & \text{otherwise,} \end{cases} \quad (16)$$

and the $M = 4$ target is $z_A + \sum_{i=1}^3 \delta_i + \tilde{\delta}_Q$. We verified 151,660 token-string overlaps, no duplicate token strings, and 151,658 shared tokens with identical ids. Only `<think>` and `</think>` require remapping; five anchor-only special tokens receive zero Qwen3 shift, and Qwen3-only special tokens are dropped. This full-vocabulary overlap projection is narrower than text-span likelihood alignment for general cross-tokenizer OPD. Because cross-tokenizer endpoint logits are not placed on the anchor vocabulary, no paired endpoint control is defined.

Adding the projected shift raises the Math macro from 33.50 to 36.82 and the five-suite macro from 26.19 to 28.04. All three Math benchmarks improve, with the largest gain on AIME 2025, while Science/IF does not improve. For reference, the Qwen3-DAPO-449 model card (<https://huggingface.co/pb09204048/Qwen3-8B-DAPO-iter449-disable-thinking>) reports AIME-24/AIME-25 pass@1 of 73.75%/67.08% under its non-thinking sampling protocol ($n = 8$, maximum response length 30,000), compared with 25.00%/17.90% for Qwen3-8B under the same protocol.

Table 13: Projected four-teacher extension at step 100, with the $M = 3$ shift composite for reference. Greedy pass@1, in percent.

Teachers	Target	AMC 2023	MATH-500	AIME 2025	Math macro	GPQA-D	IF-Eval	Sci/IF macro	Five-suite macro
$M = 3$	Δ -MOPD	38.00 ± 1.61	31.84 ± 0.99	30.67 ± 2.50	33.50	17.27 ± 1.48	13.16 ± 1.50	15.22	26.19
$M = 4$	Projected Δ -MOPD	40.50 ± 1.92	32.64 ± 1.19	37.33 ± 2.63	36.82	16.77 ± 1.76	12.94 ± 1.42	14.85	28.04