DELTA-MOPD: CROSS-TOKENIZER MULTI-TEACHER ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-teacher on-policy distillation (MOPD) integrates the capabilities of multiple teachers into a single student model through supervision on student-generated trajectories. However, existing MOPD research has largely focused on models with a shared tokenizer, limiting the integration of powerful teachers with different tokenizers. This raises a broader question: can multiple teachers with different tokenizers jointly guide the same student trajectories through parallel signal fusion? We introduce Delta-MOPD, to the best of our knowledge the first on-policy distillation framework to align and fuse policy shifts, the changes between reference and enhanced teacher checkpoints, across multiple teachers with different tokenizers. Building on prior cross-tokenizer alignment, we extend token matching to the student's Top-k candidates, enabling cross-tokenizer on-policy distillation while retaining most of their probability mass. However, we find that policy shifts based on log-probability differences can produce large signals in low-probability regions, where supervision is less reliable than in high-probability regions. We address this with probability smoothing, which, in Qwen3-4B single-teacher experiments, achieves higher average scores with only 11.4–28.7% of the absolute shift magnitude observed without smoothing. For multi-teacher fusion, we separate the directions and magnitudes of these shifts to account for differences in signal scale, and combine them according to teacher reliability. Controlled experiments focusing on signal fusion further demonstrate its effectiveness over established baselines. Across five mathematical reasoning benchmarks, Delta-MOPD improves macro-average accuracy over Qwen3-1.7B, Qwen3-4B, and Qwen3-8B by 6.86, 2.37, and 2.24 percentage points, respectively, outperforming conventional MOPD baselines and alternative signal-fusion methods. The resulting Qwen3-8B student achieves a macro-average accuracy of 61.85%, surpassing the strongest teacher, Phi-4-reasoning-plus, at 61.66%.

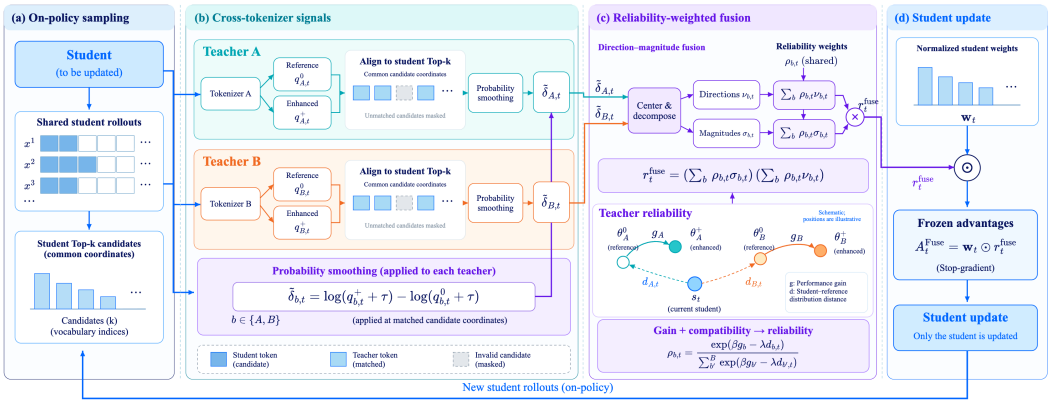


Figure 1: Overview of Delta-MOPD. Teacher checkpoint pairs score shared student rollouts; strict token alignment and probability smoothing construct policy shifts in the student Top-k space. Reliability-weighted fusion combines their directions and magnitudes into a student-weighted update. The vector geometry is schematic.

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.

1 INTRODUCTION

On-policy distillation (OPD) trains a student model with teacher supervision on its own generated trajectories (Agarwal et al., 2024). Multi-teacher OPD (MOPD) extends this approach to integrate the capabilities of multiple teachers into a single student (Ma et al., 2026). However, existing MOPD research has largely focused on teachers and students with a shared tokenizer (Gao et al., 2026; Sun et al., 2026a). This restriction narrows the choice of teachers and limits the integration of powerful models with different tokenizers.

Beyond tokenizer compatibility, MOPD must also determine how multiple teachers supervise student-generated trajectories. For example, MOPD uses prompt-level routing to assign each trajectory to a corresponding teacher (Ma et al., 2026). We explore a complementary question: can teachers with different tokenizers jointly guide the same student trajectories through parallel signal fusion? We investigate this question using policy shifts, the changes in teacher log probabilities between reference and enhanced checkpoints, as the supervision signal, following Direct-OPD (Feng et al., 2026).

Combining teacher policy shifts requires aligning them to a common set of student candidates despite differences in tokenization. However, we find that log-probability differences can produce large shifts in low-probability regions, where supervision is less reliable than in high-probability regions. Across teachers, shift magnitudes can also reflect differences in optimization intensity and need not be proportional to their performance gains.

We introduce Delta-MOPD, to the best of our knowledge the first on-policy distillation framework to align and fuse policy shifts from multiple teachers across tokenizers (Figure 1). We extend the exact-span alignment of SimpleOPD (He et al., 2026) to the student's Top-k candidates, placing teacher shifts in a shared student candidate space. Probability smoothing attenuates shifts in low-probability regions while largely preserving those in high-probability regions. We then separate each teacher's shifts into directions and magnitudes and combine them according to teacher reliability. Together, these components enable multiple teachers with different tokenizers to jointly supervise the same student trajectories within a single training stage.

Across five mathematical reasoning benchmarks, Delta-MOPD improves macro-average accuracy over Qwen3-1.7B, Qwen3-4B, and Qwen3-8B by 6.86, 2.37, and 2.24 percentage points, respectively, outperforming conventional MOPD baselines and alternative signal-fusion methods. Single-teacher experiments on Qwen3-4B show that probability smoothing achieves higher average scores with only 11.4–28.7% of the absolute shift magnitude observed without smoothing. Controlled experiments focusing on signal fusion further demonstrate its effectiveness over established baselines. The resulting Qwen3-8B student reaches a macro-average accuracy of 61.85%, surpassing the strongest teacher, Phi-4-reasoning-plus, at 61.66%.

Our contributions are threefold:

- We introduce Delta-MOPD, a cross-tokenizer multi-teacher on-policy distillation framework. By aligning policy shifts over student Top-k candidates, it enables teachers with different tokenizers to jointly supervise the same student trajectories within a single training stage.
- We develop probability smoothing to attenuate weakly supported policy shifts. In Qwen3-4B single-teacher experiments, it achieves higher average scores with only 11.4–28.7% of the absolute shift magnitude observed without smoothing.
- We develop reliability-weighted fusion that separates shift directions and magnitudes, enabling Qwen3-8B to benefit from weaker-teacher supervision and surpass the strongest teacher in five-benchmark macro-average accuracy (61.85% vs. 61.66%).

2 RELATED WORK

Multi-Teacher On-Policy Distillation. Multi-teacher on-policy distillation integrates the capabilities of multiple teachers through supervision on student-generated trajectories. Existing work studies how to allocate this supervision: MOPD (Ma et al., 2026) routes each prompt to a corresponding

teacher, while Open-MOPD (Gao et al., 2026) addresses capability imbalance through optimization-budget allocation. Beyond prompt-level routing, H-OPD (Yin et al., 2026) introduces confidence-aware token-level arbitration between vision-language and text-only teachers, and TrustMOPD (Sun et al., 2026a) uses calibrated teacher-reference displacement magnitudes to weight teacher supervision at each student state. Direct-OPD (Feng et al., 2026) transfers policy shifts between reference and enhanced teacher checkpoints rather than directly matching final teacher distributions. Building on this formulation, we study cross-tokenizer MOPD through parallel fusion of teacher policy shifts, separating their directions and magnitudes before combining them.

Cross-Tokenizer Distillation. Different tokenizers prevent direct comparison of teacher and student token distributions. Cross-tokenizer distillation addresses this mismatch through distribution matching or shared textual units: ULD (Boizard et al., 2025) uses optimal transport, while ALM (Minixhofer et al., 2025) matches likelihoods over aligned text chunks. For on-policy learning, SimCT (Sun et al., 2026b) recovers supervision through shared multi-token continuations, whereas SimpleOPD (He et al., 2026) uses exact one-to-one text-span alignment along student-generated trajectories. CompassOPD (Gu et al., 2026) further studies cross-family transfer using within-family likelihood shifts after text-space alignment. We extend exact one-to-one alignment to the student's Top-k candidates, providing a shared candidate space for multi-teacher policy-shift fusion.

3 METHOD

Delta-MOPD combines supervision from teachers with different tokenizers on shared student-generated trajectories (Figure 1). Building on Direct-OPD (Feng et al., 2026), it aligns teacher policy shifts over the student's Top-k candidates, smooths low-probability signals, and fuses the aligned shifts into a joint training signal.

3.1 POLICY-SHIFT SUPERVISION

At training iteration m, the student samples a response $y = (y_1, \ldots, y_T) \sim \pi_{\theta_m}(\cdot \mid x)$ for a prompt x. At each generated position t, the prefix $h_t = (x, y_{<t})$ defines the student distribution $p_t(v) = \pi_{\theta_m}(v \mid h_t)$. We collect its k most probable tokens into the candidate set $\mathcal{C}_t = \mathrm{TopK}_k(p_t)$.

For a teacher b, first consider a shared token space with the student. Let $q^0_{b,t}$ and $q^+_{b,t}$ denote the next-token distributions of its reference and enhanced checkpoints, conditioned on h_t. Following Direct-OPD (Feng et al., 2026), we define the policy shift as their log-probability difference:

$$\delta_{b,t}(v) = \log q^+_{b,t}(v) - \log q^0_{b,t}(v), \qquad v \in \mathcal{C}_t. \tag{1}$$

A positive shift indicates that the enhanced checkpoint assigns a higher probability to the candidate than the reference checkpoint, while a negative shift indicates a lower probability. These shifts describe how the teacher's predictions change through optimization and serve as supervision on the student's candidate tokens.

3.2 CROSS-TOKENIZER SIGNAL CONSTRUCTION

We extend the sampled-token alignment of SimpleOPD (He et al., 2026) to the student's Top-k candidates. We first align student and teacher tokens that cover the same byte span in the generated response, and denote the teacher position corresponding to student position t by $i_b(t)$.

At each aligned position, we match a student candidate to a teacher token when they produce the same byte continuation. Let $B_{s,t}(v)$ and $B_{b,i_b(t)}(u)$ denote these continuations for student candidate v and teacher token u, respectively. The candidate correspondence is

$$\mathcal{M}^{(k)}_{b,t} = \left\{ (u, v) \in \mathcal{V}_b \times \mathcal{C}_t \;\middle|\; B_{b,i_b(t)}(u) = B_{s,t}(v) \right\}. \tag{2}$$

We retain one-to-one matches, using $\mu_b(v)$ for the teacher token matched to student candidate v and $\mathcal{C}_{b,t}$ for the matched candidate set. Unaligned positions receive no supervision from teacher b. Appendix A provides the alignment details.

We note that a candidate can receive very low probabilities under both teacher checkpoints yet produce a large log-probability difference. Compared with shifts in high-probability regions, these shifts can provide less reliable supervision. To reduce their influence, we add the same smoothing constant τ to both probabilities before computing their log-ratio:

$$\widetilde{\delta}_{b,t}(v) = \log\Big(q^{+}_{b,t}(\mu_b(v)) + \tau\Big) - \log\big(q^{0}_{b,t}(\mu_b(v)) + \tau\big)\,, \qquad v \in \mathcal{C}_{b,t}. \tag{3}$$

Here, $\tau > 0$ controls the smoothing strength. Smoothing preserves each shift's sign while reducing its magnitude, with little effect when both probabilities greatly exceed τ.

3.3 Reliability-Weighted Multi-Teacher Fusion

Policy-shift magnitudes can reflect optimization intensity and need not be proportional to teacher performance gains. Directly averaging these shifts gives greater influence to larger-magnitude signals. We therefore separate direction from magnitude and fuse them using teacher reliability weights.

Direction and magnitude. Let $\bar{p}_{b,t}$ denote the student distribution normalized over teacher b's matched candidates $\mathcal{C}_{b,t}$. We subtract the student-weighted mean shift to retain relative changes in candidate preferences, then decompose the centered signal into a magnitude $\sigma_{b,t}$ and a normalized direction $\nu_{b,t}$:

$$\begin{aligned} r_{b,t}(v) &= \widetilde{\delta}_{b,t}(v) - \mathbb{E}_{u\sim\bar{p}_{b,t}}[\widetilde{\delta}_{b,t}(u)], \\ \sigma_{b,t} &= \sqrt{\mathbb{E}_{u\sim\bar{p}_{b,t}}[r_{b,t}(u)^2]}, \qquad \nu_{b,t}(v) = r_{b,t}(v)/\sigma_{b,t}. \end{aligned} \tag{4}$$

Here, $\sigma_{b,t}$ measures the student-weighted variation in the centered shift, while $\nu_{b,t}$ expresses its direction of relative preference changes over candidate tokens at a normalized scale. Appendix C provides a first-order interpretation of the centering step.

Teacher reliability. We weight each teacher using its performance gain g_b and the distance $d_{b,t}$ between its reference distribution and the student distribution at the current prefix. Teachers with larger gains and smaller distances receive higher weights:

$$\rho_{b,t} = \mathrm{softmax}_b(\beta g_b - \lambda d_{b,t})\,, \tag{5}$$

The softmax is normalized over teachers with valid supervision at the current position. The gain g_b is defined on the separate AMC23 calibration dataset and remains fixed throughout training, whereas $d_{b,t}$ varies with the student distribution and prefix. The parameters β and λ control the influence of the two terms. Appendix B defines the student probability weights, teacher gain, and student–reference distance.

Fusion. For candidates shared by all active teachers, we average teacher directions and magnitudes separately using the reliability weights, then recombine them to construct advantages:

$$A^{\mathrm{Fuse}}_t(v) = w_t(v)\left(\sum_b \rho_{b,t}\sigma_{b,t}\right)\left(\sum_b \rho_{b,t}\nu_{b,t}(v)\right). \tag{6}$$

Here, w_t is the student distribution normalized over the union of matched candidate sets. The two weighted averages combine teacher magnitudes and directions, respectively, and their product forms the fused policy shift. For candidates supported by only a subset of teachers, the reliability weights are renormalized over that subset. When only one teacher is available, we use its centered policy shift directly. Multiplying this shift by $w_t(v)$ gives the advantage for candidate v. These advantages are treated as fixed supervision for updating the student policy with Direct-OPD.

4 Experiments

4.1 Experimental Setup

Students and teachers. We use Qwen3 (1.7B, 4B, and 8B) (Yang et al., 2025) as students. The main multi-teacher experiments combine two teachers across all three student sizes: DeepSeek-R1-Distill-Qwen-1.5B → JustRL-DeepSeek-1.5B (JustRL) (DeepSeek-AI, 2025; He et al., 2025a) and

Table 1: Alignment control with Qwen3-1.7B and JustRL. Scores are AIME mean@32 (%).

Model / Method	AIME24	AIME25	Avg.
Teacher checkpoint pair			
DeepSeek-R1-Distill-Qwen-1.5B	27.50	20.80	24.15
JustRL-DeepSeek-1.5B	53.30	37.50	45.40
Student and distilled models			
Qwen3-1.7B	49.17	34.17	41.67
Native Direct-OPD	59.58	43.02	51.30
SimpleOPD	51.56	41.25	46.41
Direct-OPD (strict alignment)	57.81	42.29	50.05

Table 2: Alignment coverage (%). Mass is the mean retained share of student Top-16 probability.

Metric	Phi-4	Qwen3.5	GLM
Vocabulary	65.39	86.77	82.49
Sampled	91.11	97.48	91.88
Top-16	87.27	92.24	88.75
Mass	90.38	96.62	91.19

Phi-4-reasoning → Phi-4-reasoning-plus (Phi-4) (Abdin et al., 2025), with reference checkpoints listed before enhanced checkpoints. For the controlled fusion experiment in Section 5.3, we pair JustRL with OpenMath-Nemotron-1.5B → QuestA-Nemotron-1.5B (QuestA) (Moshkov et al., 2025; Li et al., 2026), using Qwen3-1.7B as the student. We follow Direct-OPD (Feng et al., 2026) for the JustRL and QuestA settings, despite non-identical teacher and student tokenizers (Appendix E).

Training data and optimization. Following Direct-OPD (Feng et al., 2026), we train on 105,055 mathematical prompts from Skywork-OR1 Math (He et al., 2025b). Students are trained for 300 optimizer steps in FP32 with a learning rate of 10^{-6} and batches of 64 prompts, sampling four responses per prompt. Sequential distillation uses 300 steps per teacher, totaling 600 steps. Training prompts and responses are capped at 1,024 and 2,048 tokens, respectively. Supervision uses the student's Top-16 candidates at each response position, with cross-tokenizer alignment following Section 3.2. Student updates include reference-KL regularization. For Delta-MOPD, we choose $\tau = 10^{-3}$, $\beta = 10$, and $\lambda = 1$ heuristically, without substantive hyperparameter tuning, and keep these defaults fixed across student sizes. Section 5.5 examines sensitivity to τ and β.

Evaluation. We evaluate on AIME 2024, AIME 2025, HMMT-Feb, HMMT-Nov, and OlympiadBench (He et al., 2024). We report average accuracy using 32 sampled responses per problem for AIME and HMMT (mean@32) and eight for OlympiadBench (mean@8). Decoding uses a temperature of 0.7, top-p of 0.95, and a maximum response length of 31,744 tokens. Avg. is the unweighted mean across all five benchmarks, or across AIME 2024 and AIME 2025 for AIME-only experiments. AMC23 serves as a separate calibration dataset for defining the teacher performance gain g_b (Appendix B) and is excluded from these evaluation averages. The main distillation comparisons report averages over three runs with different random seeds. Unless otherwise specified, all experiments follow the settings above.

4.2 Effectiveness of Strict Cross-Tokenizer Alignment

Despite incomplete vocabulary overlap, strict one-to-one alignment retains 87.27–92.24% of Top-16 candidates and 90.38–96.62% of their probability mass across the Phi-4, Qwen3.5, and GLM tokenizers (Table 2). Coverage is measured on 1,024 responses generated by Qwen3-1.7B from 256 Skywork-OR1 Math prompts, with four responses per prompt. Unaligned positions contribute zero to the reported probability mass.

To test whether the retained supervision remains effective for training, we compare native Direct-OPD with Direct-OPD under strict alignment using Qwen3-1.7B and JustRL. In this setting, we simulate strict one-to-one mapping by masking unmatched positions and candidates while keeping all other settings fixed. We also include SimpleOPD as a sampled-token alignment baseline.

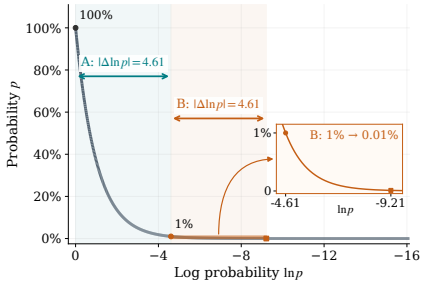
Direct-OPD under strict alignment achieves an average AIME accuracy of 50.05%, improving over the original student by 8.38 percentage points and remaining within 1.25 points of native Direct-OPD (Table 1). It also outperforms SimpleOPD by 3.64 points. These results show that strict one-to-one mapping preserves most of the performance gain from native distillation.

4.3 Effect of Probability Smoothing

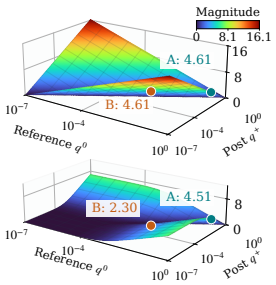
Strict token alignment does not ensure reliable supervision when both teacher checkpoints assign very low probabilities to a candidate. Figure 2 illustrates how probability smoothing reduces the magnitude of such shifts while largely preserving shifts in high-probability regions.

Quantity	A	B
Reference q^0	100%	1%
Post q^+	1%	0.01%
Change	100-fold decrease	
Original magnitude $\lvert\delta\rvert = \lvert\ln \frac{q^+}{q^0}\rvert$	A = B **4.61**	
After smoothing $\lvert\tilde{\delta}\rvert = \lvert\ln \frac{q^+ + \varepsilon}{q^0 + \varepsilon}\rvert$	**4.51**	**2.30**
Magnitude retained	**97.95%**	**50.00%**

(a) Equal relative changes



(b) Log-probability and probability changes



(c) Effect of probability smoothing

Figure 2: Illustration of probability smoothing ($\tau = 10^{-3}$). (a) Equal relative probability changes at different probability levels. (b) Probability changes on logarithmic and linear scales. (c) Original (top) and smoothed (bottom) policy-shift magnitudes. Magnitudes are in nats.

Table 3 shows that probability smoothing improves average scores with Phi-4 by 1.27, 0.50, and 0.44 percentage points on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. With JustRL, smoothing raises the Qwen3-4B average from 54.88% to 59.34%, surpassing the original student's 57.32%. Qwen3-8B also shows a smaller improvement, from 60.99% to 61.09%.

However, probability smoothing reduces performance for Qwen3-1.7B with JustRL. Our analysis suggests that JustRL provides abundant policy-shift signals in low-probability regions, and the 1.7B student's learning relies substantially on these signals. Smoothing also suppresses some valuable supervision in these regions, which may account for the observed decline. This decline is not observed with the larger Qwen3-4B and Qwen3-8B students.

4.4 Multi-Teacher Distillation Results

Using JustRL and Phi-4, we compare Delta-MOPD with single-teacher distillation, with and without probability smoothing, and three multi-teacher baselines (Table 3). MOPD Routing adapts MOPD (Ma et al., 2026) by selecting one teacher uniformly at random for each student response. Sequential distillation trains with the two teachers in successive stages, evaluating both orders. Linear Fusion averages their raw policy shifts with equal weights, without separating directions and magnitudes. For MOPD Routing and Linear Fusion, we use the available teacher's signal when only one teacher provides valid supervision.

Delta-MOPD achieves the highest average scores across all three student sizes (Table 3), improving over the original Qwen3-1.7B, Qwen3-4B, and Qwen3-8B students by 6.86, 2.37, and 2.24 percentage points, respectively. It exceeds the strongest baseline at each size by 0.21, 0.35, and 0.76 points: sequential distillation in the Phi-4-to-JustRL order on 1.7B, and single-teacher distillation with JustRL and probability smoothing on 4B and 8B. It also outperforms both sequential orderings across all three sizes using 300 rather than 600 optimizer steps. Notably, Delta-MOPD combines supervision from teachers of different strengths, including JustRL, which scores below the original Qwen3-8B student, and improves Qwen3-8B to 61.85%, surpassing the strongest teacher, Phi-4-reasoning-plus (61.66%).

5 Analysis and Ablation Studies

5.1 Component Ablation

On Qwen3-4B, we successively add probability smoothing and multi-teacher fusion to Direct-OPD with JustRL (Table 4). Smoothing improves the five-benchmark average by 4.46 percentage points,

Table 3: Single- and multi-teacher distillation across three student sizes.

Model / Method	AIME24	AIME25	HMMT Feb	HMMT Nov	Olympiad Bench	Avg.
Teacher models						
DeepSeek-R1-Distill-Qwen-1.5B	27.50	20.80	14.48	10.83	40.39	22.80
JustRL-DeepSeek-1.5B	53.30	37.50	22.29	23.75	45.42	36.45
Phi-4-reasoning	70.73	61.25	43.33	51.67	56.57	56.71
Phi-4-reasoning-plus	75.62	69.58	51.15	54.69	57.25	61.66
Qwen3-1.7B						
Original student	49.17	34.17	23.02	26.25	49.63	36.45
Single teacher: JustRL	**59.58**	43.02	24.06	33.02	51.72	42.28
+ probability smoothing	53.33	38.02	22.88	29.66	50.98	38.97
Single teacher: Phi-4	51.35	38.65	24.06	26.88	50.09	38.21
+ probability smoothing	53.75	39.27	24.90	29.06	50.41	39.48
MOPD Routing	58.96	43.75	**25.83**	33.75	51.45	42.75
Sequential (JustRL → Phi-4)	58.54	42.81	24.69	34.06	51.59	42.34
Sequential (Phi-4 → JustRL)	59.48	45.21	25.31	33.33	52.15	43.10
Linear Fusion	**59.58**	43.33	25.31	34.06	51.72	42.80
Delta-MOPD	59.17	**45.42**	25.21	**34.58**	**52.19**	**43.31**
Qwen3-4B						
Original student	70.00	67.50	41.88	51.15	56.10	57.32
Single teacher: JustRL	72.81	65.21	35.21	46.67	54.51	54.88
+ probability smoothing	76.46	70.10	40.83	**53.75**	55.56	59.34
Single teacher: Phi-4	72.29	65.94	42.29	50.00	56.32	57.37
+ probability smoothing	73.65	66.35	43.13	49.90	**56.34**	57.87
MOPD Routing	75.31	**71.15**	40.62	48.12	53.97	57.84
Sequential (JustRL → Phi-4)	72.19	65.42	38.12	47.19	54.14	55.41
Sequential (Phi-4 → JustRL)	75.21	67.40	38.75	49.38	55.75	57.30
Linear Fusion	77.50	68.02	41.88	51.67	55.93	59.00
Delta-MOPD	**78.23**	69.06	**43.33**	52.92	54.90	**59.69**
Qwen3-8B						
Original student	75.00	66.67	43.54	55.31	57.53	59.61
Single teacher: JustRL	78.96	70.00	42.08	55.52	58.40	60.99
+ probability smoothing	78.75	70.83	42.40	55.26	58.23	61.09
Single teacher: Phi-4	76.25	66.88	44.38	55.21	57.46	60.03
+ probability smoothing	74.58	68.23	**45.83**	**55.83**	57.90	60.47
MOPD Routing	76.88	71.04	41.25	54.37	57.83	60.27
Sequential (JustRL → Phi-4)	76.88	71.77	43.12	54.48	56.44	60.54
Sequential (Phi-4 → JustRL)	75.42	68.96	40.52	53.23	57.23	59.07
Linear Fusion	77.19	70.42	43.02	54.17	58.42	60.64
Delta-MOPD	**79.13**	**72.29**	43.96	55.31	**58.57**	**61.85**

Table 4: Component-wise results on Qwen3-4B. Avg. covers five benchmarks (%).

Configuration	Avg.
Base student	57.32
Direct-OPD	54.88
+ probability smoothing	59.34
+ multi-teacher fusion	**59.69**

Table 5: Smoothing parameter sensitivity on Qwen3-4B. Avg. covers AIME24/25 (%).

τ	AIME24	AIME25	Avg.
No smoothing	72.81	65.21	69.01
10^{-4}	75.83	67.50	71.67
3×10^{-4}	77.19	67.81	72.50
10^{-3}	76.46	**70.10**	**73.28**
3×10^{-3}	**77.40**	68.85	73.13

and incorporating Phi-4 through reliability-weighted fusion yields a further 0.35-point gain. The complete framework achieves 59.69%, exceeding the original student by 2.37 points.

5.2 Training Dynamics of Probability Smoothing

Figure 3 compares separate Qwen3-4B training runs with and without probability smoothing ($\tau = 10^{-3}$), using JustRL or Phi-4 as the teacher.

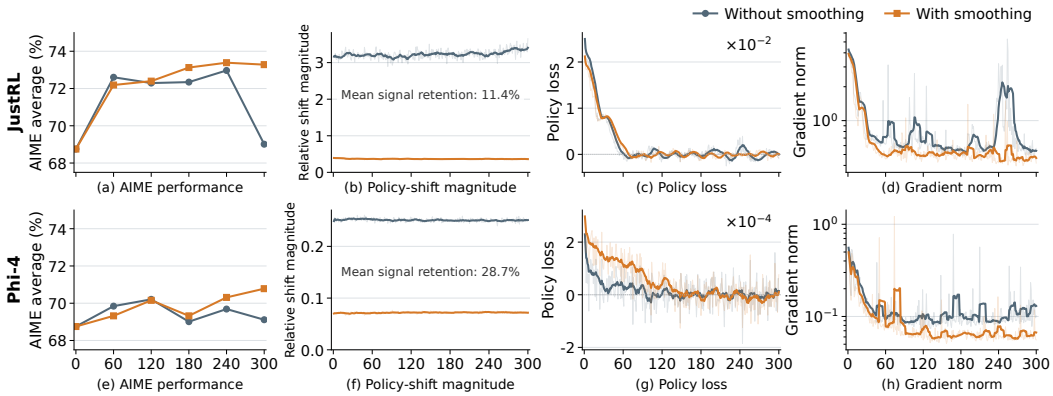


Figure 3: Training dynamics of probability smoothing on Qwen3-4B with JustRL (top) and Phi-4 (bottom). Accuracy is averaged over AIME24/25. Percentages compare absolute shift magnitudes across separate runs.

Probability smoothing attenuates unreliable policy shifts but can also suppress useful learning signals. Consistent with this trade-off, smoothing yields smaller initial accuracy gains but higher average scores at the final checkpoint. With JustRL, it also maintains late-stage accuracy while training without smoothing exhibits a sharp decline. These gains are achieved with absolute shift magnitudes only 11.4% and 28.7% of those observed without smoothing for JustRL and Phi-4, respectively.

5.3 Controlled Analysis of Multi-Teacher Fusion

To examine signal fusion with reduced tokenizer differences, we use JustRL and QuestA as teachers for Qwen3-1.7B, following the configuration in Section 4.1. We compare Delta-MOPD with single-teacher distillation, Linear Fusion, and MOPD Routing under the same initialization, training data, and 300-step budget (Figure 4). Probability smoothing is disabled for all methods to focus the comparison on signal integration.

Figure 4(a) shows that multi-teacher distillation achieves higher late-stage accuracy than either single-teacher baseline. Delta-MOPD achieves the highest accuracy from step 120 onward and retains its lead at the final checkpoint, outperforming both Linear Fusion and MOPD Routing. These results support the effectiveness of our fusion method even without probability smoothing.

Beyond accuracy, panels (b–d) characterize the distribution and scale of the training signals. These diagnostics help interpret the behavior of different methods, but do not independently determine downstream performance. Panel (b) measures the fraction of valid student Top-k probability mass assigned to candidates with positive final policy shifts:

$$M_+ = \frac{\sum_{t,v} m_{t,v} p^S_{t,v} \mathbf{1}[\Delta_{t,v} > 0]}{\sum_{t,v} m_{t,v} p^S_{t,v}}. \tag{7}$$

Here, $p^S_{t,v}$ is the student probability of candidate v at response position t, $m_{t,v}$ indicates valid supervision, and $\Delta_{t,v}$ is the final policy shift used by the corresponding method. The sums run over response positions and Top-k candidates in a training batch.

Delta-MOPD maintains a larger positive-shift probability mass throughout training, reaching approximately 0.69 in the later stages, compared with roughly 0.55 for the baselines (panel b). This indicates that positive supervision covers a larger share of the student's candidate probability mass. Its advantages have a lower standard deviation (panel c), while its gradient norm decreases early and then remains comparable to QuestA and Linear Fusion (panel d).

5.4 Teacher Activity and Policy-Shift Magnitudes

Both teachers are active at 81.08–82.90% of response positions across student sizes (Table 6). JustRL accounts for 83.49–83.89% of the total absolute magnitude of smoothed policy shifts before

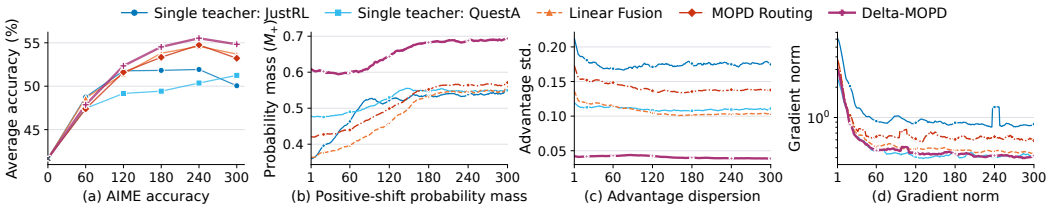


Figure 4: Performance and training dynamics on Qwen3-1.7B without probability smoothing. Accuracy averages AIME24/25 (mean@32); all horizontal axes show optimizer steps. Panels (b–d) use 11-step centered moving averages. Metric details are in Appendix F.

Table 6: Teacher activity and magnitude shares of smoothed policy shifts before fusion (%). Values are averaged over 300 training steps and three seeds.

	Active positions				Shift magnitude share	
Student	Both	Phi-4 only	JustRL only	Neither	Phi-4	JustRL
Qwen3-1.7B	81.08	2.17	6.42	10.33	16.11	83.89
Qwen3-4B	82.90	2.18	6.75	8.17	16.26	83.74
Qwen3-8B	82.40	1.85	6.81	8.94	16.51	83.49

fusion, an imbalance that may reflect differences in teacher distributions and performance gains during optimization. Appendix D provides the metric definitions and aggregation procedure.

5.5 Hyperparameter Sensitivity

We vary τ and β separately on Qwen3-4B with other settings fixed, evaluating at step 300. With JustRL, all four tested values of τ improve average scores by 2.66–4.27 percentage points over training without smoothing (Table 5). The default $\tau = 10^{-3}$ achieves the highest score of 73.28%, with similar performance at 3×10^{-3}. For β, with $\lambda = 1$ and $\tau = 10^{-3}$ fixed, the five-benchmark average varies by only 0.53 points, indicating limited sensitivity over the tested range (Table 7). The default $\beta = 10$ achieves 59.69%, close to the best score of 59.75% at $\beta = 1$.

Table 7: Sensitivity to the teacher-gain weight β on Qwen3-4B. The default is $\beta = 10$.

β	AIME24	AIME25	HMMT Feb	HMMT Nov	OlympiadBench	Avg.
0	75.52	68.73	42.08	52.50	57.31	59.22
0.1	76.67	69.48	43.02	51.35	**57.64**	59.63
1	75.31	**70.52**	**43.44**	52.08	57.38	**59.75**
10	**78.23**	69.06	43.33	52.92	54.90	59.69
15	76.88	68.54	42.40	**53.65**	55.58	59.41

6 Conclusion

We presented Delta-MOPD, which combines Top-k alignment, probability smoothing, and reliability-weighted fusion for cross-tokenizer multi-teacher on-policy distillation. Across five mathematical reasoning benchmarks and three student sizes, it outperforms the compared baselines, with Qwen3-8B surpassing the strongest teacher in macro-average accuracy. Future work will extend our two-teacher mathematical reasoning experiments to larger teacher pools and broader tasks.

AI use statement

We used generative AI tools, including OpenAI Codex, to assist with algorithm implementation, methodological and experimental-design discussions, and interpretation of results. These tools also supported literature searches, manuscript organization, English-language editing, and the preparation of framework-figure drafts. The authors retain control over methodological and editorial deci-

sions and take responsibility for reviewing and validating AI-assisted work and for the accuracy and integrity of the reported methods, results, and claims.

Ethics statement

This work studies multi-teacher distillation using existing language-model checkpoints and mathematical reasoning datasets. Distillation can transfer undesirable behaviors and biases alongside useful capabilities, and improvements on mathematical benchmarks do not establish the safety of the resulting models in other applications. Downstream use should respect the licenses and usage conditions of the underlying models and datasets and include safety evaluation appropriate to the intended application.

Reproducibility statement

Section 3 specifies the policy-shift construction and fusion method, while Section 4.1 describes the teacher and student models, training data, optimization settings, and evaluation protocol. Appendix A details cross-tokenizer alignment, and Appendices B and C provide the fusion formulation and derivation of the centering interpretation. Section 5 reports ablations and sensitivity analyses; Appendices D and F define the training diagnostics and their aggregation.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025. URL https://arxiv.org/abs/2504.21318.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2306.13649. URL https://arxiv.org/abs/2306.13649.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: The universal logit distillation loss for LLMs. *Transactions on Machine Learning Research*, 2025. doi: 10.48550/arXiv.2402.12030. URL https://arxiv.org/abs/2402.12030.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Shiyuan Feng, Huan ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation, 2026. URL https://arxiv.org/abs/2607.05394.

Huan-ang Gao, Haohan Chi, Yong Yan, Shiyuan Feng, Hanlin Wu, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Open-MOPD: Diagnosing and fixing capability imbalance in multi-teacher on-policy distillation, 2026. URL https://arxiv.org/abs/2608.19098.

Naibin Gu, Qingyi Si, Chenxu Yang, Chuanyu Qin, Junhao Zhou, Peng Fu, Zheng Lin, and Weiping Wang. CompassOPD: Cross-family on-policy distillation via within-family likelihood shifts, 2026. URL https://arxiv.org/abs/2609.10154.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Haonan He, Haodi Lei, Yun Luo, Haoran Zhang, Shunkai Zhang, Yizhuo Li, Shengji Tang, Zhilin Wang, Runzhe Zhan, Lei Bai, Ganqu Cui, Fangchen Yu, Yafu Li, Peng Ye, Ning Ding, and Yu Cheng. SimpleOPD: Simple tokenizer-agnostic on-policy distillation for long-context reasoning, 2026. URL https://arxiv.org/abs/2608.14277.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork Open Reasoner 1 technical report, 2025b. URL https://arxiv.org/abs/2505.22312.

Jiazheng Li, Hongzhou Lin, Hong Lu, Kaiyue Wen, Zaiwen Yang, Jiaxuan Gao, Yi Wu, and Jingzhao Zhang. QuestA: Expanding reasoning capacity in LLMs via question augmentation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3MifB0f7qR.

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. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training, 2026. URL https://arxiv.org/abs/2606.30406.

Benjamin Minixhofer, Ivan Vulić, and Edoardo Maria Ponti. Universal cross-tokenizer distillation via approximate likelihood matching, 2025. URL https://arxiv.org/abs/2503.20083.

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

Jie Sun, Mao Zheng, Mingyang Song, Zeyuan Liu, Gengsheng Li, Houcheng Jiang, Yilin Cheng, Bichuan Feng, Yuchen Cai, Junfeng Fang, and Xiang Wang. Distill what you trust: Reliability-aware multi-teacher on-policy distillation, 2026a. URL https://arxiv.org/abs/2609.23697.

Jie Sun, Mao Zheng, Mingyang Song, Qiyong Zhong, Yilin Cheng, Bichuan Feng, Pengfei Liu, Junfeng Fang, and Xiang Wang. SimCT: Recovering lost supervision for cross-tokenizer on-policy distillation, 2026b. URL https://arxiv.org/abs/2605.07711.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation, 2026. URL https://arxiv.org/abs/2607.02592.

A TOP-k TOKEN ALIGNMENT

Each teacher scores the student-generated response using its own tokenizer. We first align response positions, then map the student's Top-k candidates to teacher tokens with matching byte continuations.

Response-position alignment. Student and teacher tokens are represented by their byte spans in the same generated response. A student position t is aligned with teacher position $i_b(t)$ when the corresponding tokens cover the same span. Unaligned positions receive no supervision from that teacher.

Candidate mapping. At an aligned position, we match each student candidate $v \in \mathcal{C}_t$ to a teacher token with the same byte continuation under the corresponding prefix, as defined in Eq. (2). We retain one-to-one matches, denoting the teacher token by $\mu_b(v)$ and the matched student candidate set by $\mathcal{C}_{b,t}$. Candidates without a valid match do not receive supervision from teacher b.

Probability normalization. Student weights are normalized over the relevant matched candidate set, as specified in Appendix B. Teacher probabilities retain their full-vocabulary normalization when computing policy shifts and applying probability smoothing.

B ADDITIONAL FUSION DETAILS

Student probability weights. Let $\mathcal{U}_t = \bigcup_b \mathcal{C}_{b,t}$ be the union of matched candidate sets. The student weights used for fusion and for each teacher's direction–magnitude decomposition are

$$w_t(v) = \frac{p_t(v)}{\sum_{u \in \mathcal{U}_t} p_t(u)}, \qquad \bar{p}_{b,t}(v) = \frac{p_t(v)}{\sum_{u \in \mathcal{C}_{b,t}} p_t(u)}. \tag{8}$$

The two distributions are normalized over $\mathcal{U}_t$ and $\mathcal{C}_{b,t}$, respectively.

Teacher performance gain. We define teacher performance gain on AMC23, separately from the five final evaluation benchmarks. Let S_b^0 and S_b^+ denote the reference and enhanced teacher accuracies on AMC23, expressed as fractions under the same evaluation protocol. The gain is

$$g_b = \log \frac{1 - S_b^0 + \epsilon}{1 - S_b^+ + \epsilon}, \tag{9}$$

where $\epsilon > 0$ stabilizes the ratio. This log error-rate reduction is fixed before student training.

Student–reference distance. The distance between the student and teacher reference is

$$d_{b,t} = \sum_{v \in \mathcal{C}_{b,t}} \bar{p}_{b,t}(v) \left[\log \bar{p}_{b,t}(v) - \log q^0_{b,t}(\mu_b(v)) \right]. \tag{10}$$

The reference probabilities retain their full-vocabulary normalization. Thus, $d_{b,t}$ reflects both distribution mismatch and the reference probability mass assigned to the matched candidates.

C FIRST-ORDER INTERPRETATION OF CENTERING

Local softmax response. For teacher b at position t, let $z_t(v)$ be the student logits restricted to the matched candidate set $\mathcal{C}_{b,t}$, with softmax distribution $\bar{p}_{b,t}$. Applying the fixed teacher shift $\widetilde{\delta}_{b,t}$ as a small logit perturbation gives

$$\bar{p}^{(\eta)}_{b,t}(v) = \frac{\exp\Big(z_t(v) + \eta \widetilde{\delta}_{b,t}(v)\Big)}{\sum_{u \in \mathcal{C}_{b,t}} \exp\Big(z_t(u) + \eta \widetilde{\delta}_{b,t}(u)\Big)}. \tag{11}$$

Differentiating at $\eta = 0$ gives

$$\left.\frac{\partial \bar{p}_{b,t}^{(\eta)}(v)}{\partial \eta}\right|_{\eta=0} = \bar{p}_{b,t}(v)\left[\widetilde{\delta}_{b,t}(v) - \sum_{u\in\mathcal{C}_{b,t}} \bar{p}_{b,t}(u)\widetilde{\delta}_{b,t}(u)\right] = \bar{p}_{b,t}(v) r_{b,t}(v). \tag{12}$$

Here, $r_{b,t}$ is the centered shift in Eq. (4). A candidate's probability initially increases when its shift exceeds the student-weighted mean and decreases when it falls below that mean. Thus, centering identifies relative preference changes that determine the first-order probability response. Since $\sum_v \bar{p}_{b,t}(v) r_{b,t}(v) = 0$, these changes preserve total probability to first order.

Additive invariance. Adding the same constant κ to every candidate's shift leaves the centered signal unchanged:

$$\left(\widetilde{\delta}_{b,t}(v) + \kappa\right) - \mathbb{E}_{u\sim\bar{p}_{b,t}}\left[\widetilde{\delta}_{b,t}(u) + \kappa\right] = r_{b,t}(v). \tag{13}$$

Centering therefore removes shared additive offsets while preserving differences between candidate shifts, consistent with softmax invariance to a common logit offset.

D Teacher Activity and Shift-Magnitude Metrics

Teacher activity. Teacher activity is measured at generated response positions using the branch-level availability recorded during training. Each activity fraction is the proportion of all response positions in the corresponding state listed in Table 6.

Shift magnitude share. For each teacher, we sum $|\widetilde{\delta}_{b,t}(v)|$ over valid response-position–candidate pairs after probability smoothing and before centering and fusion, without student-probability weighting. Dividing this sum by the total for both teachers gives the teacher's share of the smoothed input magnitude.

Aggregation. We compute both metrics separately at each training step, average the resulting proportions equally over 300 steps, and then average across three random seeds. Table 6 reports these means rather than proportions computed from pooled training tokens.

E Tokenizer Configuration Differences

The inspected Qwen3-1.7B, JustRL, and QuestA tokenizers share the same 151,643 base token-to-ID entries and 151,387 ordered BPE merges, but differ in added-token and BOS/EOS configurations (Table 8). Within each teacher, the reference and enhanced checkpoints use identical tokenizer configurations.

Table 8: Tokenizer differences in the inspected releases. "Multi-token" indicates no single-token ID.

Property	Qwen3-1.7B	JustRL	QuestA
Added-token count	26	22	22
`<think>` ID	151667	151648	Multi-token
`</think>` ID	151668	151649	Multi-token
Token at ID 151648	`<|box_start|>`	`<think>`	`<|box_start|>`
Configured BOS ID	None	151646	None
Configured EOS ID	151645	151643	151645

Tokenizer sources. The inspected model revisions are listed below.

- Qwen3-1.7B:
  `https://huggingface.co/Qwen/Qwen3-1.7B/tree/70d244cc86ccca08cf5af4e1e306ecf908b1ad5e`
- DeepSeek-R1-Distill-Qwen-1.5B:
  `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B/tree/ad9f0ae0864d7fbcd1cd905e3c6c5b069cc8b562`

- JustRL-DeepSeek-1.5B:
  https://huggingface.co/hbx/JustRL-DeepSeek-1.5B/tree/0637e4096c789c67f9eecbe8355e0bdeddede1c2
- OpenMath-Nemotron-1.5B:
  https://huggingface.co/nvidia/OpenMath-Nemotron-1.5B/tree/59280da0e5508b98f4ae90e8886442f28ed561cb
- QuestA-Nemotron-1.5B:
  https://huggingface.co/foreverlasting1202/QuestA-Nemotron-1.5B/tree/5c0f7ec8aec547b31af6e125ea57e30622612a15

F Training Diagnostics for Controlled Fusion

Positive-shift probability mass. Equation (7) measures the fraction of valid student Top-k probability mass assigned to candidates with positive final policy shifts. It is computed on each method's own rollouts and valid candidate support. The methods apply different transformations to their shifts, and centering can change shift signs even before student parameters change. Thus, M_+ reflects signal construction as well as changes in the student distribution.

Advantage dispersion and gradient norm. Figure 4(c) reports the standard deviation of the advantage tensor used to update the student policy; panel (d) reports the student policy gradient norm. These diagnostics characterize the distribution of supervision and the gradient scale, and should be interpreted alongside accuracy.