# DIVERGENCE-ROUTED ON-POLICY DISTILLATION FOR LANGUAGE MODELS

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

## ABSTRACT

On-policy distillation (OPD) transfers knowledge from a teacher model to a student along student-generated trajectories, typically by minimizing the reverse Kullback–Leibler (RKL). However, RKL can provide weak corrective signals for teacher-supported tokens to which the student assigns very low probability. This raises two questions: which positions are most consequential for the behavioral gap between the student and teacher, and what additional supervision is needed at these positions? Through cross-sampling, we find that divergence-based routing more effectively pinpoints positions contributing to the teacher–student performance gap than entropy-based. Further analysis reveals teacher-mode under-coverage at high Jensen–Shannon (JS) divergence positions, where standard OPD increasing the large teacher–student probability gaps. Motivated by this findings, we propose **Divergence-Routed On-Policy Distillation (DrOPD)**. DrOPD retains RKL supervision at all positions, and selectively adds forward Kullback–Leibler (FKL) at positions with the high JS divergence. Experiments across six math benchmarks show that DrOPD improves over standard OPD and entropy-based routing. DrOPD also reduces the severe under-coverage cases by 37.9% relative to OPD, supporting targeted FKL as an effective complement to OPD.

## 1 INTRODUCTION

On-policy distillation (OPD) provides an effective way to transfer knowledge from a stronger teacher to a student by providing supervision along student-generated trajectories (Agarwal et al., 2023; Gu et al., 2024; Lu & Lab, 2025), thereby reducing the mismatch between offline distillation supervision and the student's inference-time state distribution. Standard OPD is commonly implemented by minimizing the reverse Kullback–Leibler (RKL) (Kullback & Leibler, 1951) between the student and teacher next-token distributions at every response position. However, RKL is weighted by the student probability, and can provide weak corrective signals for teacher-supported tokens that already receive very low student probability. As a result, some teacher modes may remain under-covered despite improvements in overall alignment.

Recent work has explored selective token-level supervision, often using the student or teacher entropy to identify important positions (Jin et al., 2026; Zhang et al., 2026; Hou et al., 2026a; Xu et al., 2026; Ko et al., 2026; Liu et al., 2026). Entropy, however, measures uncertainty within a single model, but uncertainty does not necessarily indicate a consequential difference between student and teacher. For example, both models may be uncertain while assigning similar probabilities to the available candidates. This motivates two questions: *which positions are most consequential for the teacher–student performance gap, and what additional supervision is needed at these positions?*

We use controlled *cross-sampling* (Meng et al., 2026) to identify routing signals that effectively locate positions contributing to the teacher–student performance gap. The student serves as the default generation policy, while the teacher supplies tokens only at positions selected by a routing criterion. We compare Jensen–Shannon (JS) divergence (Lin, 1991), RKL, and forward Kullback–Leibler (FKL) against student entropy, teacher entropy, and random routing. The results reveal that divergence-based strategies achieve similar intervention efficiency and outperform others. On AMC23, divergence-based strategies reach teacher performance with teacher intervention at fewer than 10% of generated positions. These results identify teacher–student disagreement as an effective signal for locating positions where additional guidance matters most.



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.

We then examine the probability distributions at high-JS positions and find pronounced teacher-mode under-coverage: on the same contexts, the student assigns substantially lower probabilities than the teacher to the recorded tokens. Standard OPD training further exacerbates severe under-coverage at these positions. This behavior is consistent with the gradient structure of RKL: for a candidate with fixed positive teacher probability, the RKL logit gradient vanishes as the student probability approaches zero, whereas FKL retains a teacher-directed corrective signal. These observations motivate adding FKL supervision at high-disagreement positions to help recover teacher-supported tokens that RKL struggles to correct.

Based on this diagnosis, we propose **Divergence-Routed On-Policy Distillation (DrOPD)**. DrOPD retains RKL supervision at all response positions and adds FKL supervision to the fraction $\rho$ of positions with the highest teacher–student JS divergence. This design preserves the standard OPD learning signal while directing complementary mode-covering supervision toward the positions identified by our diagnosis. Experiments across six mathematical reasoning benchmarks show that DrOPD improves mean Avg@8 over standard OPD by 1.84% for the 1.7B–4B student–teacher pairs. It also outperforms entropy-based routing methods, supporting disagreement as a more effective guide for allocating additional supervision. On fixed high-JS contexts, standard OPD increases the fraction of positions with severe probability gaps from 3.83% in the base student to 4.48%. DrOPD reduces this fraction to 2.79%, a relative reduction of 37.9% compared with OPD. These results connect the performance gains to improved recovery from teacher-mode under-coverage.

Our contributions are threefold:

- **Identifying where additional supervision matters.** Controlled cross-sampling shows that teacher–student divergence, RKL, FKL and JS divergence, locates behaviorally important positions more effectively than single-model entropy.
- **Diagnosing severe teacher-mode under-coverage.** Fixed-context analysis reveals severe probability gaps at high-JS positions that become more frequent after standard OPD, motivating complementary FKL supervision through gradient analysis.
- **Divergence-Routed On-Policy Distillation.** We propose DrOPD, which combines all-position RKL with selectively allocated FKL, improving reasoning performance across multiple settings and and reducing severe under-coverage.

## 2 Diagnosing Token Importance in On-Policy Distillation

We first use controlled cross-sampling to evaluate signals for identifying positions where teacher intervention improves downstream performance, then examine teacher-mode under-coverage at high-disagreement positions and its connection to KL gradients. Given a prompt $x$ and prefix $y_{<t}$, let $p_t$ and $q_t$ denote the student and teacher next-token distributions, respectively, **renormalized over the union of their top-16 candidate sets**. We compare single-model uncertainty, measured by student entropy $H(p_t)$ and teacher entropy $H(q_t)$, with teacher–student disagreement, measured by RKL $D_{\mathrm{KL}}(p_t\|q_t)$, FKL $D_{\mathrm{KL}}(q_t\|p_t)$, and JS divergence: $D_{\mathrm{JS}}(p_t, q_t) = \frac{1}{2}D_{\mathrm{KL}}(p_t\|m_t) + \frac{1}{2}D_{\mathrm{KL}}(q_t\|m_t)$, where $m_t = \frac{1}{2}(p_t + q_t)$.

### 2.1 Controlled Cross-Sampling

Following Meng et al. (2026), we use controlled cross-sampling to assess how effectively routing signals identify positions that benefit from teacher intervention. At each decoding step, giving a routing score $r_t$ and threshold $\tau$, we define $S_t = \mathbb{I}[r_t > \tau]$ and sample

$$y_t \sim \begin{cases} q_t, & S_t = 1, \\ p_t, & S_t = 0. \end{cases} \tag{1}$$

The sampled token is appended to the prefix. We compare routing based on JS divergence, RKL, FKL, student entropy, and teacher entropy, together with a random baseline. For each stategy, we independently sweep the threshold $\tau$; for random routing, we manually change the intervention probability. Each setting is evaluated on 40 AMC23(Li et al., 2024) problems, with eight responses

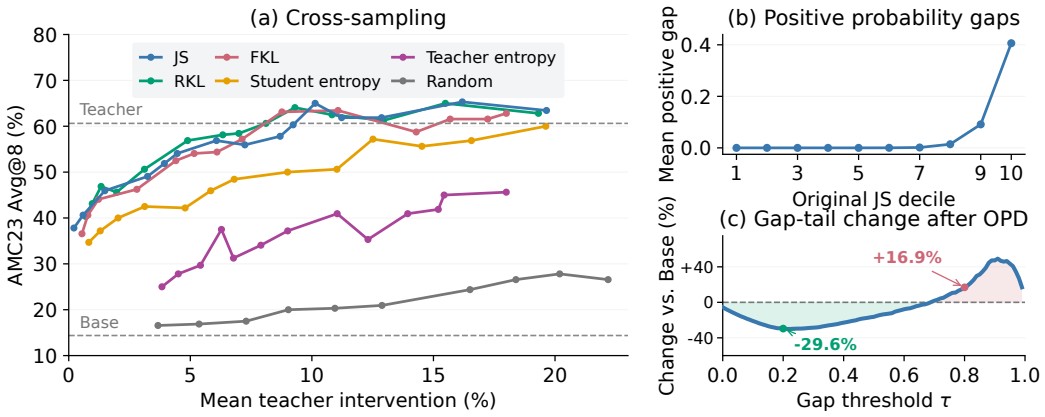


Figure 1: **Cross-sampling and teacher-mode under-coverage at high-JS positions.** (a) AMC23 accuracy versus teacher-intervention rate; dashed lines indicate standalone Teacher and Base performance. (b) Mean positive probability gap across original JS deciles, computed over recorded tokens with positive gaps. (c) Relative change from Base to OPD in the fraction of positions with gaps at or above each threshold, evaluated on fixed original JS $> 0.06$ positions.

per problem. We measure the actual teacher-intervention ratio as:

$$\eta = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i} S_t^{(i)}, \qquad N = 320, \tag{2}$$

where $T_i$ is the length of response $i$. We compare Avg@8 as a function of $\eta$: higher accuracy at comparable intervention ratios indicates that a routing signal more effectively identifies positions where teacher guidance improves task performance.

### 2.2 Divergence Identifies Behaviorally Critical Positions

Figure 1a shows that divergence-based routing uses teacher intervention more efficiently than others. JS divergence, RKL, and FKL achieve similar performance, reaching near-teacher accuracy with intervention at fewer than 10%. For example, JS routing achieves 60.31% Avg@8 with 9.23% intervention, close to the teacher's 60.63%. In contrast, student- and teacher-entropy routing achieve only 50.00% and 37.19%, respectively. These results suggest that divergence is an effective signal for locating behaviorally critical positions, while entropy alone is less informative because a model can be uncertain even when the student and teacher make similar predictions.

We use JS divergence for the remaining analysis and the routing mechanism in Section 3 because it is symmetric and bounded. Having identified high-disagreement positions is valuable, we next examine their probability distributions to understand what additional supervision they need.

### 2.3 From High Disagreement to Teacher-Mode Under-Coverage

**Measuring under-coverage on fixed contexts.** We examine the probability distributions at high-JS positions using 320 Base–Teacher cross-sampled responses. To quantify teacher-mode under-coverage, we define the positive teacher–student probability gap as

$$\delta_t = \big[q_t(v) - p_t(v)\big]_+, \qquad [a]_+ = \max(a, 0), \tag{3}$$

where $v$ is the recorded token at position $t$. A larger $\delta_t$ indicates that the student assigns less probability to candidate $v$ than the teacher does.

Figure 1b shows that under-coverage is more pronounced at high-JS positions. For the base student, the mean positive gap increases from 0.014 in the eighth JS decile to 0.406 in the highest. Thus, positions identified as valuable intervention sites also exhibit substantial deficits in the probability assigned to teacher-supported tokens. Standard OPD reduces many probability gaps but exacerbates

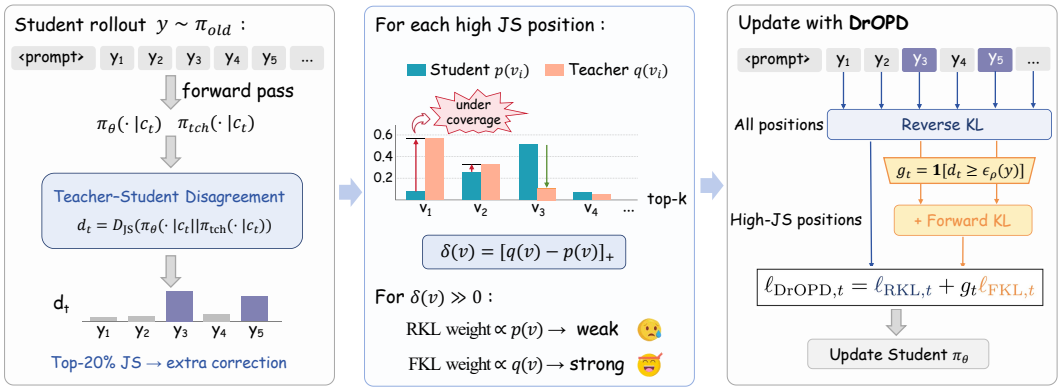


Figure 2: **Overview of DrOPD.** (a) We compute teacher–student JS divergence at each rollout position and select the top-20% with the highest disagreement. (b) High-JS positions can exhibit under-coverage of teacher-supported modes, for which RKL provides weak corrective signals. (c) DrOPD retains RKL supervision at all positions and adds FKL supervision at the high-JS positions.

severe under-coverage. Figure 1c shows the relative change from Base to OPD in the fraction of fixed high-JS positions with $\delta_t \geq \tau$, where $\tau$ is the gap threshold. At $\tau = 0.2$, this fraction decreases by 29.6%. At $\tau = 0.8$, however, it increases from 3.83% to 4.48%, a relative increase of 16.9%. Both comparisons use the same positions with original JS $> 0.06$.

**Understanding under-coverage through gradient analysis.** The persistence of severe under-coverage is consistent with the gradient structure of RKL. For a student distribution $p$ parameterized by logits $z(v)$ and a fixed teacher distribution $q$, the logit gradients are

$$\frac{\partial D_{\mathrm{KL}}(p\|q)}{\partial z(v)} = p(v)\left(\log\frac{p(v)}{q(v)} - D_{\mathrm{KL}}(p\|q)\right), \tag{4}$$

$$\frac{\partial D_{\mathrm{KL}}(q\|p)}{\partial z(v)} = p(v) - q(v). \tag{5}$$

For fixed $q(v) > 0$, the RKL gradient vanishes as $p(v) \to 0$, weakening the direct corrective signal for severely under-covered candidates. The FKL gradient instead approaches $-q(v)$, so gradient descent continues to increase the student's logit for these candidates. This complementary gradient behavior motivates adding FKL supervision at high-disagreement positions.

## 3 Divergence-Routed On-Policy Distillation

Motivated by the above findings, we propose DrOPD, which retains the RKL throughout student-generated trajectories and adds FKL at selected high-JS positions. Figure 2 illustrates the method.

### 3.1 Disagreement Scoring and Position Selection

Set a student policy $\pi_{\mathrm{old}}$ to generate student responses. At position $t$, let $c_t$ denote the generated prefix. We form the union candidate set from the student's and teacher's top-$k$ candidates, and renormalize the two distributions over the union candidate: :

$$\mathcal{C}_t = \mathrm{TopK}_k(p_t) \cup \mathrm{TopK}_k(q_t), \quad \widetilde{p}_t(v) = \frac{p_t(v)}{\sum_{u\in\mathcal{C}_t} p_t(u)}, \quad \widetilde{q}_t(v) = \frac{q_t(v)}{\sum_{u\in\mathcal{C}_t} q_t(u)}. \tag{6}$$

This union retains both student-preferred candidates and teacher-supported candidates. We use $k = 16$ and score disagreement at each position by $d_t = D_{\mathrm{JS}}(\widetilde{p}_t, \widetilde{q}_t)$. Let $y$ denote the collection of student responses in the current batch, and let $\epsilon_\rho(y)$ be the $(1-\rho)$-quantile of disagreement scores over its valid response positions. We select high-disagreement positions using

$$g_t = \mathbb{1}[d_t \geq \epsilon_\rho(y)]\,. \tag{7}$$

We set $\rho = 0.2$, allocating additional supervision to the top 20% highest-disagreement positions.

Table 1: **Mathematical reasoning results (Avg@8, %).** Both panels select and average over all six benchmarks. Results marked with † are reproduced by us.

| Method | AIME24 | AIME25 | AMC23 | Minerva | Olympiad | MATH500 | Mean(6) |
|---|---|---|---|---|---|---|---|
| ***Baselines*** | | | | | | | |
| 1.7B-Base | 0.42 | 1.25 | 14.38 | 6.34 | 8.18 | 23.88 | 9.07 |
| 4B-Base | 10.00 | 5.42 | 30.00 | 6.48 | 24.78 | 40.38 | 19.51 |
| 4B-GRPO | 24.17 | 17.50 | 60.63 | 30.19 | 48.53 | 83.10 | 44.02 |
| 8B-GRPO | 25.83 | 21.67 | 68.13 | 35.29 | 51.52 | 86.00 | 48.07 |
| ***Student: 1.7B-Base Teacher: 4B-GRPO*** | | | | | | | |
| OPD | 12.92 | 8.33 | 45.31 | 21.23 | 33.03 | 68.63 | 31.57 |
| Stu-entropy | 12.08 | 7.08 | 47.19 | 24.59 | 34.55 | 70.43 | 32.65 |
| Tch-entropy | 12.92 | 7.92 | 46.25 | 23.44 | 34.77 | 71.00 | 32.72 |
| EOPD† (2026) | 11.25 | 7.50 | 45.94 | 23.07 | 34.72 | 69.93 | 32.07 |
| DrOPD | 14.17 | 11.67 | 46.56 | 24.03 | 33.88 | 70.15 | 33.41 |
| ***Student: 4B-Base Teacher: 8B-GRPO*** | | | | | | | |
| OPD | 24.58 | 18.33 | 61.88 | 33.00 | 47.92 | 82.15 | 44.64 |
| Stu-entropy | 23.33 | 19.17 | 59.38 | 32.86 | 47.61 | 82.93 | 44.21 |
| Tch-entropy | 20.83 | 19.17 | 62.19 | 33.64 | 47.57 | 82.53 | 44.32 |
| EOPD† (2026) | 22.08 | 18.75 | 62.19 | 32.67 | 48.02 | 83.15 | 44.48 |
| DrOPD | 25.00 | 19.17 | 63.44 | 32.54 | 48.07 | 83.18 | 45.23 |

### 3.2 All-Position RKL with Routed FKL Supervision

Using $\widetilde{p}_t$ and $\widetilde{q}_t$, we define the top-$k$ approximate distillation loss at position $t$ as

$$\ell_t(\theta) = \sum_{v \in \mathcal{C}_t} \widetilde{p}_t(v) \log \frac{\widetilde{p}_t(v)}{\widetilde{q}_t(v)} + g_t \sum_{v \in \mathcal{C}_t} \widetilde{q}_t(v) \log \frac{\widetilde{q}_t(v)}{\widetilde{p}_t(v)}. \tag{8}$$

The RKL preserves the standard OPD objective at every response position. When $g_t = 1$, the additional FKL supplies teacher-weighted supervision to help correct under-covered candidates.

In practice, we optimize a policy-gradient surrogate using the following distillation reward:

$$R_t(v) = -\widetilde{p}_t(v) \log \frac{\widetilde{p}_t(v)}{\widetilde{q}_t(v)} + g_t \widetilde{q}_t(v), \qquad v \in \mathcal{C}_t. \tag{9}$$

The first term supplies the standard OPD signal, while the second adds the FKL correction. We use these rewards directly as detached advantages, $A_t(v) = \mathrm{sg}[R_t(v)]$, and update the student by maximizing the PPO-style clipped surrogate:

$$\mathcal{J}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \sum_{v \in \mathcal{C}_t} \min[w_t(v;\theta) A_t(v), \mathrm{clip}(w_t(v;\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})\, A_t(v)]\,. \tag{10}$$

Here, $w_t(v;\theta)$ is the importance-sampling weight, $\mathcal{B}$ contains the valid response positions in the batch, and $\epsilon_{\mathrm{clip}}$ is the clipping parameter. The teacher remains frozen during each policy update.

## 4 Experiments

### 4.1 Experimental Settings

**Models and training.** Our primary pair uses Qwen3-1.7B-Base as the student and Qwen3-4B-Base-GRPO as the teacher. We also evaluate a Qwen3-4B-Base student with a Qwen3-8B-Base-GRPO teacher (Yang et al., 2025; Shao et al., 2024). We use verl(Sheng et al., 2024) as our training framework. Students train on DAPO-Math-17K(Yu et al., 2026a) for one epoch with four student-generated responses per prompt, with batch size of 64, and a 7,168-token response limit. DrOPD uses the union of both models' top-16 candidates and routes added FKL to the top-20% JS positions. All diagnostic and sensitivity analyses use the primary model pair.

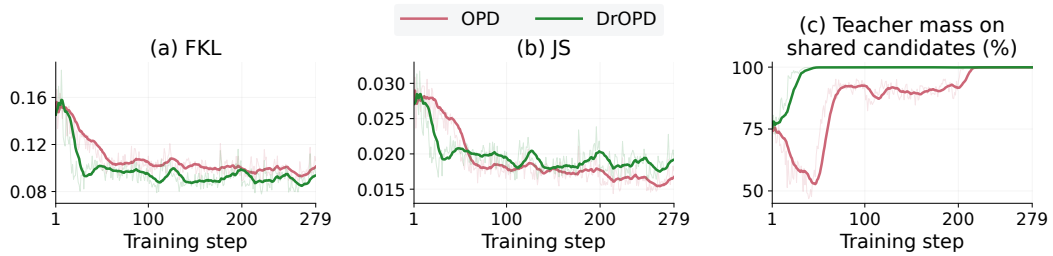


Figure 3: **Additional forward-KL accelerates early alignment.** Changes in (a) FKL and (b) JS divergence during training. (c) Full-softmax teacher mass on the top-16 intersection.

**Evaluation.** We report AIME24, AIME25 (Balunovic et al., 2025), AMC23 (Li et al., 2024), Minerva (Lewkowycz et al., 2022), Olympiad (He et al., 2024), and MATH500 (Hendrycks et al., 2021). Each problem has eight responses sampled at temperature 0.7 and top-$p = 0.95$, with a 7,168-token limit. Avg@8 measures mean response correctness.

**Compared methods.** We compare DrOPD with standard OPD and entropy-based supervision strategies. **OPD** applies RKL supervision at all response positions. **Stu-entropy** and **Tch-entropy** retain all-position RKL and add FKL at the top 20% of positions ranked by the corresponding entropy. We also reproduce **EOPD** (Jin et al., 2026), which adds FKL at positions where teacher entropy exceeds 0.8 while retaining RKL supervision at all positions.

### 4.2 Main Results

Table 1 shows that DrOPD improves over standard OPD on both model pairs. For the 1.7B–4B pair, mean Avg@8 increases from 31.57% to 33.41%, a gain of 1.84 percentage points, with improvements on all six benchmarks. For the 4B–8B pair, DrOPD raises the mean from 44.64% to 45.23% and improves five of the six benchmarks. These results show that the benefit of targeted FKL supervision extends across both student scales.

DrOPD also outperforms entropy-based routing for both model pairs. For the 1.7B–4B pair, stu-entropy routing, tch-entropy routing, and the EOPD achieve 32.65%, 32.72%, and 32.07%, respectively, compared with 33.41% for DrOPD. This comparison highlights the value of the routing criterion. Consistent with the cross-sampling results, teacher–student disagreement provides a more effective guide for allocating forward supervision than either model's uncertainty alone.

### 4.3 Training Dynamics

We examine how DrOPD changes alignment during training through FKL, JS divergence, and teacher probability mass on the intersection of the two top-16 candidate sets (Figure 3). With additional FKL supervision, DrOPD maintains lower FKL than OPD over most of training (Figure 3a). DrOPD also reduces JS divergence more rapidly during the early updates (Figure 3b), consistent with directing additional supervision to high-disagreement positions. This faster initial alignment is consistent with the earlier gains and higher peak in reasoning accuracy shown in Figure 4c, although its JS divergence becomes slightly higher than OPD's later in training. Teacher probability mass on the top-16 intersection provides a complementary view of candidate coverage (Figure 3c). This metric sums the teacher's probabilities over the top-k shared candidates retained by both models. Under DrOPD, the mass quickly approaches 100% and remains high, indicating that the student's top-16 candidates capture nearly all teacher probability mass. In contrast, OPD shows larger fluctuations and reaches a comparable level substantially later. These dynamics suggest that routed FKL helps earlier coverage of teacher-supported candidates while reducing the distributional discrepancy.

### 4.4 Under-Coverage and Learning Progress

**Recovering under-covered teacher modes.** We evaluate the same recorded tokens on the fixed high-JS contexts from Section 2.3, comparing Base, OPD, and DrOPD on the same prefixes and

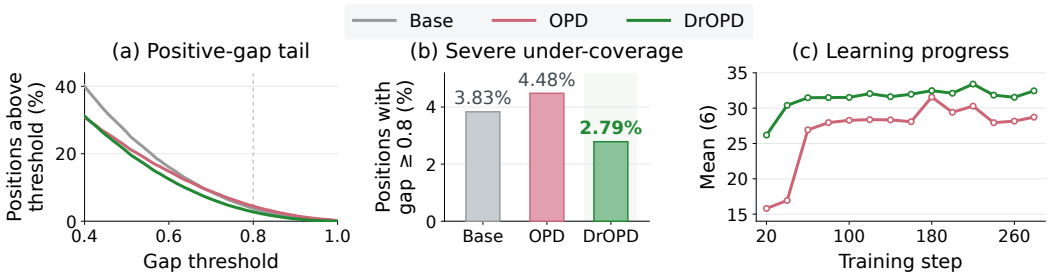


Figure 4: **Under-coverage recovery and learning progress.** (a) Fraction of fixed high-JS positions with probability gaps at or above each threshold. (b) Severe under-coverage ($\delta_t \geq 0.8$) on the same positions. (c) Mean Avg@8 across six benchmarks at successive training checkpoints.

recorded tokens. Figure 4a reports the fraction of positions with $\delta_t \geq \tau$ as the gap threshold $\tau$ varies. DrOPD consistently reduces this proportion relative to OPD for thresholds from 0.5 to 0.9, indicating improved recovery across a range of under-coverage severities. At the severe-gap threshold of 0.8 (Figure 4b), the proportion decreases from 4.48% under OPD to 2.79% under DrOPD, a relative reduction of 37.9%, and falls below the base's 3.83%. These results support the role of routed FKL in correcting teacher-supported tokens that remain severely underweighted after RKL training.

**Faster gains in reasoning accuracy.** The mean accuracy in Figure 4c shows that DrOPD also learns useful task behavior earlier. Its mean accuracy of six benchmarks reaches 30.40% at step 40, whereas OPD first exceeds 30% at step 180. DrOPD also achieves a higher peak accuracy, showing that its gains extend beyond the early training stage. Together with the training dynamics in Section 4.3, these results show that improved distributional alignment and reduced severe under-coverage accompany earlier and stronger gains in reasoning performance.

### 4.5 Parameter Sensitivity

We analysis the effects of the routing fraction $\rho$, the candidate count $k$, and candidate-set construction in Figure 5. Due to computational constraints, these experiments use four benchmarks: AIME24, AIME25, AMC23, and MATH500. Each problem samples eight responses and report mean Avg@8 across these benchmarks. All experiments use the 1.7B–4B model pair.

**Routing fraction.** Adding FKL supervision to the top 10% of high-JS positions increases mean Avg@8 from 33.80% to 34.63%, and 20% reaches the best, while larger routing fractions yield no further improvement (Figure 5a). These results favor concentrating additional supervision on a moderate fraction of high-disagreement positions rather than extending it broadly across the trajectory.

**Candidate count.** When $k$ is small, the truncated distribution can differ substantially from the full-vocabulary, limiting the accuracy of the loss approximation. Increasing $k$ improves this approximation. Figure 5b shows that performance gains saturate around $k = 16$, which we choose as the default to balance accuracy and computational cost.

**Candidate-set construction.** Using only the student's candidates can exclude teacher-supported tokens that need FKL correction, while using only the teacher's candidate may omit student-preferred tokens relevant to RKL alignment. Their union retains both groups, supporting both RKL alignment and FKL correction. As shown in Figure 5c, the union achieves the highest mean Avg@8 of 35.64%, compared with 34.79% for student-only and 35.23% for teacher-only selection.

### 4.6 Loss Direction and Supervision Allocation

Our cross-sampling results suggest that a small subset of high-disagreement positions accounts for much of the benefit of teacher intervention. This raises a natural question: *can distillation supervision also be restricted to these positions?* To investigate this, we rank positions by teacher–student

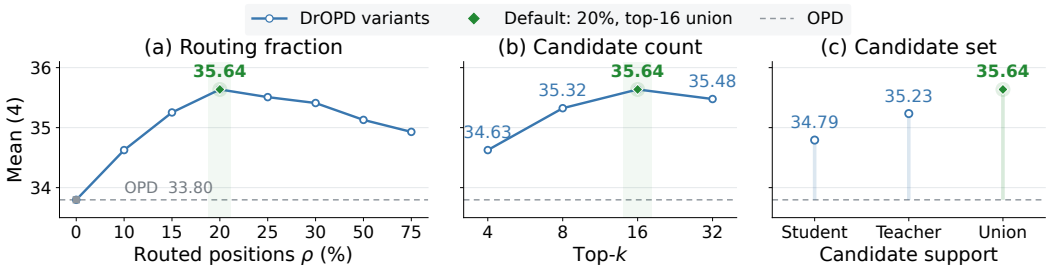


Figure 5: **Sensitivity to routing and candidate support.** Avg@8 averaged over four benchmarks for (a) the routed fraction, (b) top-$k$ per model, and (c) candidate-set construction. The default uses top-20% JS routing with a top-16 union. The dashed line denotes OPD, equivalent to $\rho = 0$.

Table 2: **Effect of loss direction and supervision allocation.**. Selected positions are the top 20% ranked by JS; remaining positions are the other 80%. "–" denotes no loss at those positions.

| Method | Remaining positions | Selected positions | Mean(4) |
|---|---|---|---|
| OPD (Full RKL) | RKL | RKL | 33.80 |
| Full FKL | FKL | FKL | 34.16 |
| Sparse RKL | – | RKL | 34.94 |
| Sparse FKL | – | FKL | 33.29 |
| DrOPD | RKL | RKL + FKL | **35.64** |

JS divergence and divide them into the bottom 80% and top 20%. We then apply different combinations of RKL and FKL to these two regions, as summarized in Table 2. All variants use the same top-16 union candidate set and are evaluated by mean Avg@8 across the four benchmarks.

Full RKL (standard OPD) and Full FKL apply the respective loss to both regions, whereas Sparse RKL and Sparse FKL supervise only the top-20% high-JS positions. DrOPD retains RKL in both regions and adds FKL only to the top 20%. Sparse RKL improves mean Avg@8 from 33.80% to 34.94%, supporting the value of concentrating supervision at high-disagreement positions. However, the same restriction reduces FKL performance from 34.16% to 33.29%. Thus, the benefit of sparse supervision depends on the loss direction. DrOPD achieves the highest mean Avg@8 of 35.64%. Together, these results favor combining broad RKL alignment with selective FKL supervision over applying either objective alone in the tested configurations.

The strong Sparse RKL also raises a further question: could two directions benefit from different position or allocations that change during training? DrOPD provides a simple effective allocation, while jointly adapting the placement and relative strength of RKL and FKL remains for future work.

## 5 Related Work

**On-policy distillation and its mechanisms.** OPD trains a student on its own generated trajectories using teacher feedback, reducing the mismatch between training contexts and those encountered at inference. GKD provides a framework for learning from student-generated sequences with different divergence objectives (Agarwal et al., 2023), while MiniLLM develops RKL-based distillation to suppress student probability in regions that the teacher considers unlikely (Gu et al., 2024). ExOPD generalizes this formulation through reward extrapolation (Yang et al., 2026). Recent studies examine when this supervision succeeds, highlighting teacher–student compatibility, teacher reliability on student prefixes, and the effects of divergence estimation (Li et al., 2026a; Fu et al., 2026; Zhu et al., 2026). Corresponding improvements include more informative state exploration and calibrated teacher guidance in Uni-OPD (Hou et al., 2026a), variance reduction through control variates in vOPD (Oh et al., 2026), and explicit treatment of omitted probability mass in Tail-Aware Top-$k$ OPD (Huang & Wei, 2026). Self-distillation constructs a teacher from the model itself: SDPO conditions the model on textual feedback (Hübotter et al., 2026), while Self-Distilled Reasoner uses verified reasoning traces as privileged context (Zhao et al., 2026). Our work focuses on a fixed

external teacher and studies how to supplement RKL at positions where teacher-supported tokens remain severely under-covered.

**Token importance and selective supervision.** Selective distillation allocates supervision according to the usefulness or reliability of individual tokens. REOPOLD combines entropy-based token sampling with reward clipping (Ko et al., 2026), and SelecTKD uses teacher verification to retain or down-weight token-level distillation losses (Huang et al., 2025). FiRe-OPD filters trajectories using teacher likelihood and then reweights tokens using teacher confidence and student uncertainty (Li et al., 2026b). TIP jointly analyzes student entropy and teacher–student divergence, identifying low-entropy, high-disagreement positions that entropy-only selection can miss (Xu et al., 2026). In self-distillation, PW-OPSD uses position-dependent weighting to account for teacher reliability along reasoning trajectories (Liu et al., 2026), while DASH uses disagreement to adapt the temporal extent of supervision (Hou et al., 2026b). Meng et al. (2026) use cross-sampling between base and RL-trained models to show that sparse high-divergence positions account for much of the performance change after RLVR. Building on this diagnostic, we compare routing signals through their downstream effects at comparable teacher-intervention ratios. We further examine the probability deficits at high-JS positions on fixed contexts.

**Divergence objectives and teacher-mode recovery.** Several methods adapt the form of supervision to address limitations of a uniform distillation objective. EOPD retains RKL and adds FKL where teacher entropy exceeds a threshold (Jin et al., 2026). AOPD retains reinforcement updates for positive-advantage tokens and replaces non-positive-advantage updates with imitation over teacher top-$k$ candidates (Jia et al., 2026). TIDE independently selects student-excess and student-deficit positions, applying bounded Hellinger corrections to the former and teacher top-$k$ supervision to recover missing probability mass at the latter (Yu et al., 2026b). TRACE applies FKL to annotated key reasoning spans, optionally applies RKL to error spans, and combines distillation with GRPO; it also supports online self-annotation (Wang et al., 2026). In privileged-context self-distillation, DASD uses entropy to choose between attraction toward and repulsion from the self-teacher (Zhang et al., 2026). DrOPD shares the motivation of recovering under-covered teacher modes, but uses a different supervision allocation: all positions retain RKL, and the highest-JS positions receive additional FKL. Our contribution connects this allocation to cross-sampling evidence of behavioral importance and fixed-context evidence of severe under-coverage, without requiring reasoning-span annotations.

## 6 Conclusion

We study where standard on-policy distillation fails to effectively transfer teacher knowledge. Controlled cross-sampling shows that teacher–student distributional disagreement identifies behaviorally critical positions more effectively than single-model uncertainty, while further analysis reveals severe teacher-mode under-coverage at many of these positions. Motivated by this diagnosis, we propose Divergence-Routed On-Policy Distillation (DrOPD), which retains RKL supervision throughout student-generated trajectories and selectively adds FKL supervision to the highest-disagreement positions. Experiments across multiple mathematical reasoning benchmarks and student–teacher configurations show that DrOPD consistently improves over standard OPD and entropy-based routing strategies. Our analyses further support targeted mode-covering supervision as an effective way to address teacher knowledge that is difficult to recover with RKL alone.

## References

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. *arXiv preprint arXiv:2306.13649*, 2023.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0679.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting On-Policy Distillation: Empirical Failure Modes and Simple Fixes. *arXiv preprint arXiv:2603.25562*, 2026. URL https://arxiv.org/abs/2603.25562v2.

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

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, 2024. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026a.

ZhiYan Hou, Xinyu Tang, Hongyan An, Jianjin Zhang, Weizhen Wang, Yunyun Han, Gengsheng Li, Xiangzhao Hao, Haiyun Guo, Wenbin Hu, Jinqiao Wang, and Yafeng Deng. DASH: Divergence-Adaptive Supervision Horizons for On-Policy Self-Distillation of Reasoning Models. *arXiv preprint arXiv:2608.06243*, 2026b. URL https://arxiv.org/abs/2608.06243v2.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. Selectkd: Selective token-weighted knowledge distillation for llms, 2025. URL https://arxiv.org/abs/2510.24021.

Huipeng Huang and Hongxin Wei. Tail-Aware Top-$k$ On-Policy Distillation. *arXiv preprint arXiv:2608.14728*, 2026. URL https://arxiv.org/abs/2608.14728v1.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation. *arXiv preprint arXiv:2601.20802*, 2026. doi: 10.48550/arXiv.2601.20802. URL https://arxiv.org/abs/2601.20802.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric On-Policy Distillation: Bridging Exploitation and Imitation at the Token Level. *arXiv preprint arXiv:2605.06387*, 2026. URL https://arxiv.org/abs/2605.06387v4.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

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

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022. URL https://arxiv.org/abs/2206.14858.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. Numina, Hugging Face dataset, 2024. URL https://huggingface.co/datasets/AI-MO/NuminaMath-CoT.

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

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, Then Reweight: Rethinking Optimization Granularity in On-Policy Distillation. *arXiv preprint arXiv:2606.02684*, 2026b. URL https://arxiv.org/abs/2606.02684v2.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *arXiv preprint arXiv:2605.21606*, 2026.

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

Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. Sparse but critical: A token-level analysis of distributional shifts in rlvr fine-tuning of llms. *arXiv preprint arXiv:2603.22446*, 2026.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. KL for a KL: On-Policy Distillation with Control Variate Baseline. *arXiv preprint arXiv:2605.07865*, 2026. URL https://arxiv.org/abs/2605.07865v1.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Jiaxuan Wang, Xuan Ouyang, Zhiyu Chen, Yulan Hu, Zheng Pan, Xin Li, and Lan-Zhe Guo. Trace: Distilling where it matters via token-routed self on-policy alignment, 2026. URL https://arxiv.org/abs/2605.10194.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token Importance in On-Policy Distillation. *arXiv preprint arXiv:2604.14084*, 2026. doi: 10.48550/arXiv.2604.14084. URL https://arxiv.org/abs/2604.14084.

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

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation. *arXiv preprint arXiv:2602.12125*, 2026. doi: 10.48550/arXiv.2602.12125. URL https://arxiv.org/abs/2602.12125.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch Matters: On-Policy Distillation Beyond Token Agreement. *arXiv preprint arXiv:2608.09836*, 2026b. URL https://arxiv.org/abs/2608.09836v1.

Hongbin Zhang, Chaozheng Wang, Kehai Chen, Youcheng Pan, Yang Xiang, Jinpeng Wang, and Min Zhang. Tailoring teaching to aptitude: Direction-adaptive self-distillation for llm reasoning. *arXiv preprint arXiv:2605.22263*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. *arXiv preprint arXiv:2601.18734*, 2026. doi: 10.48550/arXiv.2601.18734. URL https://arxiv.org/abs/2601.18734.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The Many Faces of On-Policy Distillation: Pitfalls, Mechanisms, and Fixes. *arXiv preprint arXiv:2605.11182*, 2026. doi: 10.48550/arXiv.2605.11182. URL https://arxiv.org/abs/2605.11182.