# CONTRASTIVE ON-POLICY DISTILLATION VIA DUAL-BRANCH SUPERVISION

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

## ABSTRACT

For on-policy distillation (OPD), more capable models are a natural choice as teachers. However, higher task performance does not necessarily imply greater teachability. A controlled token intervention study suggests that even a model weaker than the student can serve as an informative contrastive reference. Building on this observation, we propose Contrastive On-Policy Distillation (cOPD), which provides dual branch supervision by contrasting strong teacher, weak teacher, and student predictions at student generated prefixes. Our analysis explains how this three way comparison helps select token candidates associated with better or worse reasoning outcomes under an idealized model. We evaluate cOPD across diverse teacher student configurations spanning different training stages, parameter scales, and model families, on tasks including mathematical reasoning, instruction following, and question answering with abstention. cOPD achieves the best or top-tier performance among the evaluated baselines across all three tasks, improving overall Pass@16 on mathematical reasoning by up to 7.51% over the strongest baseline. Further analyses show that cOPD suppresses undesirable response patterns and suggest that it facilitates exploration.

## 1 INTRODUCTION

Distillation has rapidly emerged as a core technique for large language model (LLM) post-training. In particular, On-Policy Distillation (OPD) (Agarwal et al., 2024) samples trajectories from the student and obtains supervision from the teacher's next-token distributions at the resulting prefixes, providing dense, token level guidance. Recent LLMs, including Qwen3 (Yang et al., 2025), DeepSeek-V4 (Xu et al., 2026), and GLM-5 (Zeng et al., 2026), have incorporated OPD into their post-training pipelines for capability transfer and consolidation, highlighting its growing practical importance alongside supervised fine-tuning (SFT) and reinforcement learning (RL).

As OPD uses the teacher's predictive distribution as a training target, using more capable models as teachers is both a natural and widely adopted design choice (Jin et al., 2026; Yang et al., 2025; Gu et al., 2024). However, recent work shows that higher teacher accuracy does not necessarily translate into greater student improvement through OPD (Li et al., 2026b; Fu et al., 2026). A model's ability to solve tasks is therefore distinct from its teachability in distilling a particular student. Although a model with lower performance may be a less reliable target for direct distillation, its distribution might still contain useful information for improving the student.

This key observation motivates our research question in this work: **Can a model that performs worse than the student still provide beneficial supervision?** We investigate weaker model's role as an auxiliary teacher, using its predictive distribution as a contrast to that of a primary, stronger teacher. Our hypothesis is that comparing their token level preferences can help identify both beneficial token choices to reinforce and harmful ones to suppress. A token intervention study provides initial support for this hypothesis. Figure 1 shows that takeovers guided by weak and strong teacher contrasts can improve student accuracy when favoring the strong teacher and reduce it when favoring the weak teacher.

In this paper, we propose Contrastive On-Policy Distillation (cOPD). At student generated prefixes, cOPD selects token candidates whose probabilities under the student lie between those assigned by the strong and weak teachers. Through the teachers' log-probability contrasts, cOPD provides dual-branch supervision, where the positive branch reinforces selected preferences favored by the



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.

strong teacher, while the negative branch suppresses those favored by the weak teacher. Our analysis explains how comparing the strong teacher, student, and weak teacher helps select token candidates associated with better or worse reasoning outcomes under an idealized model. We evaluate cOPD across diverse teacher student configurations on mathematical reasoning, instruction following, and question answering with abstention, and find that it achieves the best or top-tier performance among the evaluated baselines. Further empirical analyses show that cOPD effectively uses weak teacher information to suppress undesirable response patterns, with qualitative evidence suggesting benefits for exploring additional valid responses.

Our specific contributions are as follows:

1. We show through controlled token takeovers that weak teachers provide complementary information for identifying beneficial and harmful token preferences.
2. We propose cOPD, which uses strong weak teacher contrasts to selectively reinforce beneficial token preferences and suppress harmful ones.
3. We provide extensive empirical evidence across three tasks and diverse teacher student configurations that cOPD achieves best or near-best performance while suppressing undesirable failure modes.

## 2 PRELIMINARY

Let $x \sim \mathcal{D}$ denote an input prompt and $y = (y_1, ..., y_L)$ a response, with $y_{<t} = (y_1, ..., y_{t-1})$. We denote the student and a teacher by $\pi_\theta$ abd $\pi_T$, with their next token distributions as $\pi(\cdot|x, y_{<t})$.

**On-Policy Distillation.** OPD trains the student policy using teacher's supervision on its own generated trajectories. Given a prompt $x$ from a training set $\mathcal{D}$, the student generates a response $y \sim \pi_\theta(\cdot|x)$. At each position $t$, the teacher provides a next-token distribution $\pi_T(\cdot|x, y_{<t})$, conditioned on the prompt and the student-generated prefix. The student is updated by minimizing a token-level divergence between the teacher and student distributions:

$$\mathcal{L}_{OPD}(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}[\frac{1}{L}\sum_{t=1}^{L} D(\pi_T(\cdot|x, y_{<t}), \pi_\theta(\cdot|x, y_{<t}))]$$

where $D$ is a divergence measure and the sampled trajectories are treated as fixed during each update.

**Distillation objectives.** Common choices for the token level divergence include forward KL, reverse KL, and Jensen-Shannon divergence (JSD) (Song & Zheng, 2026). Let $p(v) = \pi_T(v|x, y_{<t})$ and $q(v) = \pi_\theta(v|x, y_{<t})$, where $v \in \mathcal{V}$ and $\mathcal{V}$ is the full vocabulary. The Kullback-Leibler (KL) (Kullback & Leibler, 1951) divergence between two distribution $r(v)$ and $s(v)$ is defined as:

$$D_{KL}(r||s) = \sum_{v\in\mathcal{V}} r(v) \log \frac{r(v)}{s(v)}$$

*Forward KL*, $D_{FKL}(p, q) = D_{KL}(p||q)$, encourages mode-covering behavior by penalizing insufficient student probability on tokens supported by the teacher.

*Reverse KL*, $D_{RKL}(p, q) = D_{KL}(q||p)$, encourages mode-seeking behavior by penalizing student probability mass on tokens unlikely under the teacher.

*Jensen-Shannon divergence*, $D_{JSD}(p, q) = \frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m)$, where $m = \frac{1}{2}(p + q)$, provides a symmetric, bounded alternative by comparing both distributions to their mixture.

Most OPD implementations approximate the divergence using a top-$k$ subset of the vocabulary (Jin et al., 2026; Li et al., 2026b).

## 3 METHOD

### 3.1 CAN A WEAK TEACHER HELP?

OPD uses teacher supervision at student generated prefixes to reshape the student's next token predictions. We design a controlled token intervention experiment to investigate whether a weak teacher

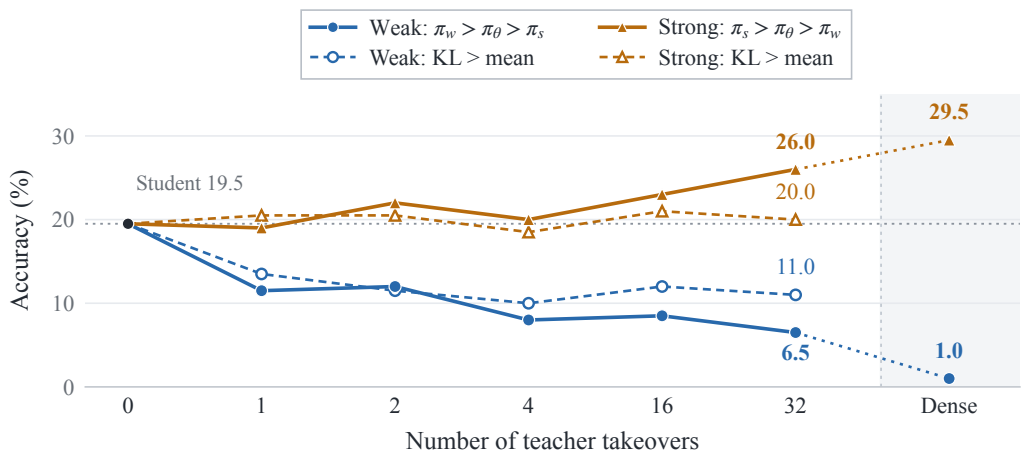


Figure 1: Teacher takeover experiment. As the takeover budget increases, the negative branch with weak teacher takeovers generally reduces student accuracy, whereas the positive branch with strong teacher takeovers generally improves it. Both effects are larger than under KL-based selection.

can serve as a contrastive reference for identifying critical positions and token distributions that affect the correctness of the student's final answer. The experiment uses temporary teacher takeovers as a simplified behavioral proxy for the effects of OPD supervision.

We use DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025) as the student $\pi_\theta$, Qwen-2.5-Math-1.5B (Yang et al., 2024) as the weak teacher $\pi_w$, and JustRL-DeepSeek-1.5B (He et al., 2025) as the strong teacher $\pi_s$. We select a fixed set of 50 problems from the DAPO-Math-17k dataset (Yu et al., 2025) for which the weak teacher pass@8 $= 0$. Each intervention setting is evaluated on this set with four random seeds. Given the current prefix $h_t = (x, y_{<t})$, the three models assign probabilities to the candidate tokens $v \in \mathrm{Top}_8(\pi(\cdot|h_t))$. We define two takeover conditions as:

$$\begin{aligned} \text{Weak teacher takeover:} \quad & \pi_{\mathrm{w}}(v \mid h_t) > \pi_\theta(v \mid h_t) > \pi_{\mathrm{s}}(v \mid h_t), \\ \text{Strong teacher takeover:} \quad & \pi_{\mathrm{s}}(v \mid h_t) > \pi_\theta(v \mid h_t) > \pi_{\mathrm{w}}(v \mid h_t), \end{aligned}$$

At the first eligible position during student generation, the corresponding teacher takeover and generates four consecutive tokens. The student then resumes generation until the next eligible position. This procedure repeats until the takeover budget is reached, and then student completes the response. We evaluate budget of $n \in \{1, 2, 4, 16, 32, \infty\}$ takeovers. We also evaluate a separate control that triggers takeovers where KL divergence exceeds the mean KL divergence.

As shown in Figure 1, contrastive based selection produces an overall increase in student accuracy with strong teacher takeovers and decrease with weak teacher takeovers as the budget grows. These effects are more significant than takeovers under KL based selection suggesting that weak teacher may help to provide additional information for identifying takeover positions and distributions that improve or degrade student accuracy. This motivates incorporating these weak strong teacher contrasts into OPD supervision to reinforce beneficial distributions and suppress harmful ones.

### 3.2 Contrastive On-Policy Distillation

Motivated by these findings, we propose Contrastive On-Policy Distillation (cOPD), which uses three way probability comparisons among the strong teacher, student, and weak teacher to construct token level supervision at student generated prefixes.

#### 3.2.1 Contrastive token mask

We first define a token level mask that selects token candidates for which the student probability lies between the weak and strong teacher probabilities, as in the motivation experiment. To exclude near ties, we require both probability gaps to exceed a margin $\delta$.

For a student generated prefix $h_t = (x, y_{<t})$, we have $\pi_s$, $\pi_w$, and $\pi_{ref}$ as the strong teacher, weak teacher, and reference student policy, respectively, where $v \in \mathrm{Top}_{16}(\pi_{ref}(\cdot|h_t))$. We define two binary masks corresponding to the positive and negative branches:

$$\begin{aligned} m_v^+ &= \mathbf{1}[\pi_s(v|h_t) - \pi_{ref}(v|h_t) > \delta \wedge \pi_{ref}(v|h_t) - \pi_w(v|h_t) > \delta], \\ m_v^- &= \mathbf{1}[\pi_w(v|h_t) - \pi_{ref}(v|h_t) > \delta \wedge \pi_{ref}(v|h_t) - \pi_s(v|h_t) > \delta], \end{aligned} \tag{1}$$

where $\mathbf{1}[\cdot]$ is the indicator function. We keep $\pi_{ref}$ fixed within each epoch and set $\pi_{ref} \leftarrow \pi_\theta$ at the end of the epoch. The positive branch selects tokens favored by the strong teacher relative to the reference student and disfavored by the weak teacher. The negative branch selects tokens with the reverse ordering. Tokens satisfying neither condition fall into the abstain case. The overall selection mask is

$$m_v = m_v^+ + m_v^-. \tag{2}$$

#### 3.2.2 CONTRASTIVE OBJECTIVE

We first quantify the relative preference of the strong and weak teachers through their log probability contrast:

$$\Delta_v = \log \pi_s(v) - \log \pi_w(v) = \log \frac{\pi_s(v)}{\pi_w(v)}. \tag{3}$$

A positive $\Delta_v$ indicates that the strong teacher prefer token $v$ over the weak teacher, while a negative value indicates the reverse. Its magnitude reflects the difference in their relative preferences.

Let $q_\theta(v) = \pi_\theta(v|h_t)$ be the current student probability. We combine the teacher contrast with the token mask to define the per-position loss:

$$\ell_{cOPD}(\theta; h_t) = -\sum_{v \in \mathrm{Top}_{16}(\pi_{ref})} m_v q_\theta(v) \tanh \frac{\Delta_v}{\tau}. \tag{4}$$

where the temperature $\tau > 0$ controls the scale of the contrast. Positive branch tokens have $\Delta_v > 0$ and encourage student probability mass on these tokens, and negative branch tokens have $\Delta_v < 0$ and penalize student probability mass. Abstaining tokens contribute no term to the loss.

### 3.3 ANALYSIS OF CONTRASTIVE SUPERVISION

To understand how a weak teacher can provide complementary information for selecting training signals, we present a theoretical analysis.

Prior work analyzes token level teacher guidance through downstream success probabilities, distinguishing beneficial, neutral, and detrimental contributions (Armandpour et al., 2026). For a fixed student generated prefix $h$, we model the idealized next token distribution of an autoregressive language policy as the following mixture:

$$p_i = \alpha H_i + \beta_i G + \gamma_i B, \quad i \in \{s, ref, w\}, \tag{5}$$

where $p_s$, $p_{ref}$, $p_w$ correspond to the strong teacher, reference student, and weak teacher, and $\alpha + \beta_i + \gamma_i = 1$. The normalized distribution $G$ and $B$ are shared across models and represent good and bad reasoning candidates, that under a fixed student continuation, their correctness probabilities are respectively above and below the reference student's expected correctness at $h$. The task neutral components $H_i$ may differ across models, such as preferring "but" over "however" when the two expressions are semantically equivalent in context. We assume that the supports of $G$, $B$, and the combined neutral region are mutually disjoint, with $0 < \alpha < 1$ and $0 < \beta_w < \beta_{ref} < \beta_s < 1 - \alpha$.

**Proposition 1** (Contrastive selection of reasoning preferences)**.** *Under the shared component model, token candidates in the reasoning components satisfy*

$$\begin{aligned} v \in \mathrm{supp}(G): &\quad p_{\mathrm{s}}(v) > p_{ref}(v) > p_{\mathrm{w}}(v), \qquad m_v^- = 0, \\ v \in \mathrm{supp}(B): &\quad p_{\mathrm{w}}(v) > p_{ref}(v) > p_{\mathrm{s}}(v), \qquad m_v^+ = 0. \end{aligned}$$

*Within the reasoning components, $m_v^+$ and $m_v^-$ serve as selective proxies for good and bad reasoning choices, respectively. The cOPD loss uses these proxies to encourage probability mass on selected good reasoning candidates and penalize mass on selected bad reasoning candidates.*

Table 1: Results on mathematical reasoning (%). Avg. is the overall Pass@16 across all problems pooled from AIME24/AIME25/AIME26/AMC23. Bold and underlined scores indicate the best and second best results.

| **Method** | AIME24 Pass@16 ↑ | AIME25 Pass@16 ↑ | AIME26 Pass@16 ↑ | AMC23 Pass@16 ↑ | Avg. Pass@16 ↑ | Omni-Math Pass@1 ↑ |
|---|---|---|---|---|---|---|
| *JustRL / Qwen2.5-Math / DeepSeek-R1-Distill* | | | | | | |
| Base | 70.00 | 46.67 | 40.00 | 87.95 | 69.36 | 16.84 |
| Strong teacher | 76.67 | 60.00 | 40.00 | 93.98 | 75.72 | 28.18 |
| Weak teacher | 23.33 | 13.33 | 10.00 | 66.27 | 39.88 | 10.35 |
| FKL | <u>76.67</u> | <u>53.33</u> | <u>43.33</u> | 91.57 | 73.99 | 12.19 |
| RKL | 73.33 | 46.67 | 40.00 | <u>93.98</u> | 72.83 | 15.14 |
| JSD | <u>76.67</u> | **56.67** | <u>43.33</u> | 92.77 | <u>75.14</u> | 14.14 |
| MAD-OPD | 73.33 | 50.00 | 40.00 | 87.95 | 70.52 | 16.66 |
| W2S-OPD | <u>76.67</u> | 46.67 | 36.67 | <u>93.98</u> | 72.83 | **25.49** |
| **cOPD (ours)** | **80.00** | **56.67** | **46.67** | **95.18** | **77.46** | <u>23.82</u> |
| *Qwen3-4B / 0.6B / 1.7B* | | | | | | |
| Base | 33.33 | 30.00 | 36.67 | 74.70 | 53.18 | 11.63 |
| Strong teacher | 46.67 | 46.67 | 53.33 | 83.13 | 65.32 | 20.60 |
| Weak teacher | 13.33 | 13.33 | 10.00 | 56.63 | 33.53 | 11.34 |
| FKL | <u>46.67</u> | 33.33 | 26.67 | <u>78.31</u> | 56.07 | <u>16.13</u> |
| RKL | 36.67 | <u>36.67</u> | 20.00 | 77.11 | 53.18 | 15.92 |
| JSD | 43.33 | 30.00 | <u>30.00</u> | <u>78.31</u> | 55.49 | 15.77 |
| MAD-OPD | 36.67 | 23.33 | 16.67 | 75.90 | 49.71 | 15.95 |
| W2S-OPD | 40.00 | <u>36.67</u> | <u>30.00</u> | **81.93** | <u>57.80</u> | 16.02 |
| **cOPD (ours)** | **63.33** | **50.00** | **36.67** | **81.93** | **65.32** | **16.41** |

*Proof.* For $v \in \mathrm{supp}(G)$, the disjoint-support assumption gives $p_{\mathrm{s}}(v) - p_{ref}(v) = (\beta_{\mathrm{s}} - \beta_{ref})G(v)$ and $p_{ref}(v) - p_{\mathrm{w}}(v) = (\beta_{ref} - \beta_{\mathrm{w}})G(v)$. Both gaps are positive, yielding $p_{\mathrm{s}}(v) > p_{ref}(v) > p_{\mathrm{w}}(v)$, while the negative branch remains inactive. Since $\gamma_i = 1 - \alpha - \beta_i$, the same argument applies to $B$ with the teacher ordering reversed, yielding $p_{\mathrm{w}}(v) > p_{ref}(v) > p_{\mathrm{s}}(v)$ and $m_v^+ = 0$. The teacher log-ratio is positive on $G$ and negative on $B$, giving the corresponding reinforcement and suppression directions in the loss.

The probability margin $\delta$ further restricts which candidates are selected without changing these directional conclusions. The analysis illustrates how a teacher that assigns less probability mass to good reasoning can still provide an informative contrastive reference. cOPD uses this additional reference to select supervision from the relative predictions of the strong and weak teachers.

## 4 Experiments

We evaluate whether cOPD improves On-Policy Distillation by using a weak teacher as a contrastive reference. Our experiments span mathematical reasoning, instruction following, and question answering with abstention. The teacher student configurations cover different post-training backgrounds, parameter scales, and model families.

### 4.1 Experiment setup

We use three model configurations, listed below in the order of strong teacher, weak teacher, and student:

1. JustRL-DeepSeek-1.5B (He et al., 2025), Qwen2.5-Math-1.5B (Yang et al., 2024), and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025). These models share Qwen2.5-1.5B as base model but differ in post-training.
2. Qwen3-4B, Qwen3-0.6B, and Qwen3-1.7B (Yang et al., 2025). These models belong to the same family but differ in parameter scale.

3. Llama-3.1-8B-Instruct, Llama-3.2-1B-Instruct, and Llama-3.2-3B-Instruct. These models span different releases and parameter scales within the Llama family.

The model configurations used for each task, together with the corresponding training and evaluation settings, are described in the following subsections.

We compare cOPD against six baselines: the initial student model (Base), OPD with forward KL (FKL), reverse KL (RKL), or Jensen–Shannon divergence (JSD), multi teacher debate OPD (MAD-OPD) (Wang et al., 2026a), and weak to strong multi treacher OPD (W2S-OPD) (Yu et al., 2026). All models are run in non-thinking mode with detailed training hyperparameters provided in Appendix A.

### 4.2 Mathematical Reasoning

**Setup and evaluation.** We use DAPO-Math-17k (Yu et al., 2025) for training with the three model configurations. We exclude the Llama configuration because we do not observe a clear performance gap between the teachers and the student on mathematical reasoning. For each configuration, we retain prompts for which the weak teacher produces no correct response and the strong teacher produces at least one correct response in eight samples. This filtering results in 6,016, 7,488 training prompts for the JustRL and Qwen3 configurations. We evaluate on AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), AIME26 (Zhang & Math-AI, 2026), and AMC23 using Pass@16, and on Omni-Math (Gao et al., 2025) using Pass@1.

**Results.** We report mathematical reasoning results in Table 1. cOPD achieve the highest overall Pass@16 among all distillation methods in both configurations, reaching $77.46\%$ with JustRL and $65.32\%$ with Qwen3. These results outperform the strongest baselines by up to 7.51 percentage points. More importantly, cOPD surpasses the strong teacher in the JustRL configuration and matches the strong teacher in the Qwen3 configuration. On Omni-Math, cOPD achieves the best student performance with Qwen3 and ranks second to W2S-OPD with JustRL.

We further analyze truncation as a failure mode in the JustRL configuration in Figure 3(a). The weak teacher exhibits frequent truncation errors and low accuracy, whereas the strong teacher achieves higher accuracy with fewer truncation errors. Strong-teacher-only OPD transfers this favorable behavior, reducing the student's truncation errors from 321 to 54–65 while improving accuracy.

Incorporating the weak teacher, however, does not consistently preserve these benefits. MAD-OPD has 318 truncation errors with no accuracy gain over the initial student. This suggests that a weak teacher whose responses are frequently truncated may provide limited useful feedback during debate, limiting student improvement. W2S-OPD matches the strong teacher's accuracy but retains more truncation errors than single-teacher OPD, suggesting that their incorporation of weak teacher may limit the suppression of truncation failures.

cOPD reduces truncation errors to 56 while achieving the highest accuracy, even surpassing the strong teacher. These results suggest that cOPD can exploit the weak teacher's complementary information to improve performance while suppressing its failure modes.

### 4.3 Instruction Following

**Setup and evaluation.** We use the multi constraint IF-RLVR (Pyatkin et al., 2025) dataset, with up to five verifiable constraints per prompt. A fixed set of 7,000 prompts is held out for testing. For each model configuration, we retain training prompts for which the weak teacher produce no correct response, and the strong teacher produces at least one correct response in eight samples, yielding 6,892 and 6,903 training prompts for Qwen3 and Llama.

**Results.** We report evaluation results in Table 2. cOPD consistently outperforms all distillation baselines in both model configurations, achieving $24.26\%$ accuracy with Qwen3, and $21.73\%$ with Llama. These results exceed the strongest baseline in each configuration, RKL for Qwen3 and W2S-OPD for llama, by $0.80$ and $1.96$ percentage points, respectively. Moreover, on Llama, cOPD surpasses the strong teacher by $3.00$ percentage points.

Table 2: Results on instruction following (IF) and question answering with abstention (QA w/ Abstention). All scores are percentages, and Abst. represents the correct abstention rate. Bold indicates the best student result in each column, and underline indicates the second best.

| | *Qwen3-4B/ 0.6B/ 1.7B* | | | | *Llama-3.1-8B/ 3.2-1B/ 3.2-3B* | | | |
|---|---|---|---|---|---|---|---|---|
| | IF | QA w/ Abstention | | | IF | QA w/ Abstention | | |
| Method | Acc.↑ | EM↑ | F1↑ | Abst.↑ | Acc.↑ | EM↑ | F1↑ | Abst.↑ |
| Base | 20.36 | 60.14 | 65.64 | 55.68 | 17.81 | 54.20 | 60.82 | 46.24 |
| Strong teacher | 27.13 | 68.35 | 74.39 | 62.32 | 18.74 | 67.07 | 71.75 | 69.50 |
| Weak teacher | 18.33 | 35.01 | 39.80 | 30.63 | 11.69 | 16.63 | 21.23 | 8.53 |
| FKL | 23.20 | <u>71.43</u> | <u>76.49</u> | <u>69.86</u> | 19.35 | 65.71 | 70.05 | 68.23 |
| RKL | <u>23.46</u> | 67.75 | 73.33 | 63.13 | 19.11 | 66.43 | 70.65 | 68.73 |
| JSD | 22.48 | 70.04 | 75.40 | 68.33 | 18.66 | <u>67.36</u> | <u>71.75</u> | <u>70.58</u> |
| MAD-OPD | 20.10 | 63.18 | 69.97 | 57.78 | 16.07 | 48.75 | 51.10 | 64.44 |
| W2S-OPD | 22.81 | 66.17 | 71.28 | 59.60 | <u>19.77</u> | 3.40† | 4.01† | 1.95† |
| **cOPD (ours)** | **24.26** | **72.09** | **77.31** | **71.59** | **21.73** | **68.97** | **73.17** | **73.04** |

† For Llama QA, W2S-OPD exhibits a catastrophic formatting error that causes verifier failures. Ignoring this error yields EM, F1, and correct abstention rate scores of 34.69, 41.27, and 13.29.

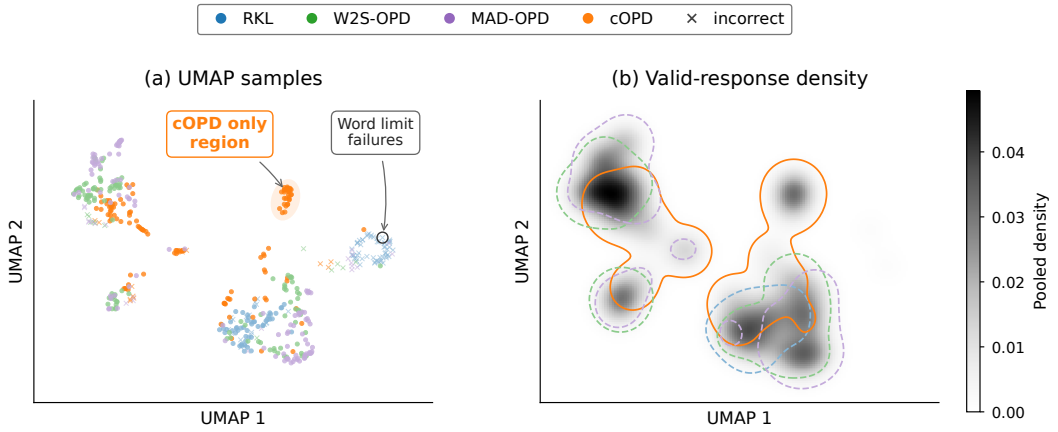


Figure 2: Instruction-following response distributions and constraint satisfaction in a shared UMAP space, where each point represents a response and crosses indicate violation. cOPD produces valid responses in additional regions while largely avoiding the prominent word-limit failure region.

Beyond the accuracy results, we further visualized responses generated using identical prompts in a shared UMAP space in Figure 2. Our method covers the major regions containing valid baseline responses. It also reach a distinct cluster of correct responses absent from the baseline samples, demonstrating its ability to explore additional regions while satisfying all instruction constraints. Meanwhile, cOPD largely avoids the prominent word limit failure region, confirming our hypothesis that cOPD can suppress harmful response patterns. We provide illustrative response excerpts in Appendix C.

### 4.4 QA WITH ABSTENTION

**Setup and evaluation.** We use SQuAD2.0 (Rajpurkar et al., 2018), which contains both answerable and unanswerable questions. Given a context and a question, the model is instructed to answer based on the context and abstain when the context does not support an answer. From the training split, we retain prompts for which the weak teacher produces no correct response and the strong teacher produces at least one correct response in eight samples. This filtering yields 8,080 and 16,359 training prompts for Qwen3 and Llama, with unanswerable questions accounting for

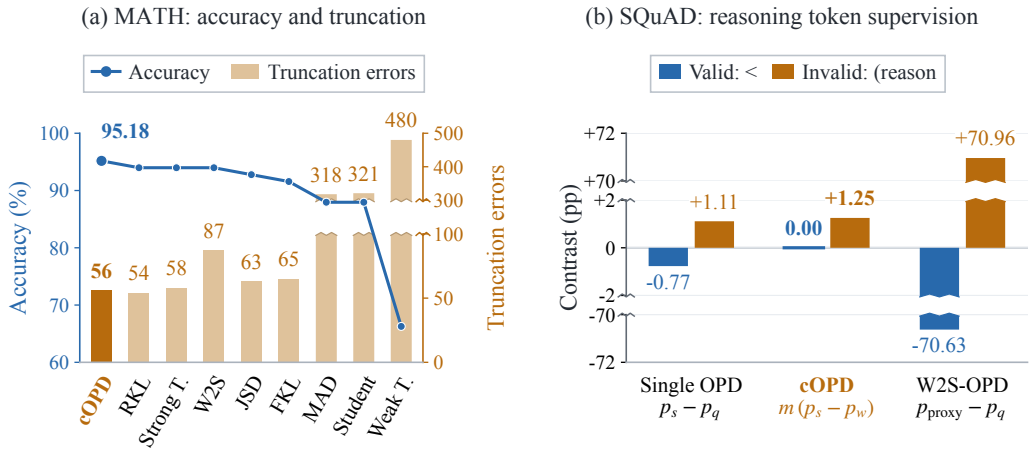


Figure 3: **Failure mode analysis.** (a) Mathematical reasoning accuracy (Pass@16) and truncation error counts. cOPD suppresses truncation failures while surpassing the strong teacher in accuracy. (b) Opening-token supervision for valid and invalid reasoning formats. W2S-OPD excessively amplifies the invalid token's relative weight, whereas cOPD's corresponding gap is negligible.

26.51% and 49.91% of the training subsets. We evaluate on the full official development split of 11,873 questions.

**Results.** We report QA with abstention results in Table 2. cOPD outperforms all distillation baselines and even the strong teacher on EM, F1, and correct abstention rate in both configurations. cOPD achieves a correct abstention rate of 71.59% on Qwen3, exceeding the strongest baseline by 1.73 percentage points and on Llama, it reaches 73.04%, a gain of 2.46 percentage points over the strongest baseline.

We further investigate the catastrophic degradation of W2S-OPD on SQuAD with the Llama configuration. We observe that the trained model incorrectly starts its reasoning with `(reason` instead of `<`, violating the required `<reasoning>` format. To examine this behavior, we sampled 128 training prompts and compute the initial student's and both teachers' distributions at the first response position. Figure 3(b) compares the teacher student probability gaps for both tokens between OPD's single teacher, W2S-OPD's proxy teacher, and the student. For cOPD, we report the masked probability contrast $m_v(p_s - p_w)$ from Equation 2 and 3 as a diagnostic supervision coefficient.

W2S-OPD excessively amplifies the relative weight of the invalid token `(reason`, raising its average probability from 1.25% under the strong teacher to 71.10% under the proxy teacher. This distorted target provides an explanation for the observed formatting failures.

In comparison, cOPD uses a margin gated mask to selectively apply contrastive supervision. Its estimated supervision coefficient for the invalid token remains small, without the extreme amplification observed in W2S-OPD. This results suggest that cOPD can use weak teacher information while remain robust to failures modes.

### 4.5 Discussion of Results

Our empirical evidence can answer the motivating research question affirmatively: **A model that performs worse than the student can provide beneficial supervision when used as a contrastive reference alongside a stronger teacher.** Across mathematical reasoning, instruction following, and question answering with abstention, cOPD achieves the best or tied-best performance among the evaluated OPD methods. These gains span different model families, parameter scales, and post training backgrounds, supporting the value of weak-teacher contrasts across the evaluated settings.

The benefits also extend to suppressing undesirable response patterns. In mathematical reasoning, cOPD achieves the highest accuracy in the analyzed configuration while substantially reducing truncation errors (Figure 3(a)). In instruction following, it largely avoids the prominent failure region

associated with word limit violations (Figure 2). On SQuAD2.0, it avoids the catastrophic formatting failure observed with W2S-OPD in the Llama configuration (Figure 3(b)). Together, these results indicate that cOPD can use weak teacher information as an counterexample to avoid weak teacher's failure modes.

Additionally, the instruction-following visualization suggests a potential benefit for exploration, that cOPD reaches a distinct region of satisfying responses absent from the baselines (Figure 2). Ablations in Appendix B show the contribution of both supervision branches and dataset selection.

# 5 Related Works

## 5.1 On-Policy Distillation

OPD trains students on their own generated trajectories using token-level teacher supervision, reducing the distribution mismatch between training and inference (Song & Zheng, 2026; Agarwal et al., 2024; Kim & Rush, 2016). Following studies have refined distillation objectives to account for teacher uncertainty (Jin et al., 2026; Li et al., 2026a) and examined how teacher–student compatibility affects knowledge transfer (Niu et al., 2026; Wang et al., 2026c; Liu et al., 2026). The reliance of standard OPD on a single teacher motivates exploring complementary supervision from multiple models.

## 5.2 Multi-Teacher OPD

Multi-teacher OPD methods broadly follow two approaches, teacher routing and teacher aggregation. Routing based methods keep single teacher supervision at the question level by assigning each student generated trajectory to a domain specific teacher (Ma et al., 2026; Gao et al., 2026). Whereas aggregation based methods combine supervision from multiple teachers for the same trajectory (Wang et al., 2026a). Our work falls within the broader scope of multi teacher OPD, while focusing on a different question: how models weaker than the student can contribute useful supervision.

## 5.3 Weaker Model Supervision

Prior work has demonstrated the value of weak supervision in other learning settings, where stronger models are trained on labels generated by weaker models (Burns et al., 2023; Yuan et al., 2026), as well as rejected responses used as negative examples in preference learning (Wang et al., 2026b).

Direct-OPD (Feng et al., 2026) uses the log probability contrast between a weak model's pre and post RL checkpoints as token level supervision for stronger students. W2S-OPD (Yu et al., 2026) constructs a proxy teacher by adding the logit contrast between positive and negative models to the student's logits. Our method supports teacher combinations beyond paired pre and post RL checkpoints, and cOPD does not construct a proxy teacher. Instead, it uses the reference student as the anchor for three way probability comparisons to select token-level supervision.

# 6 Conclusion

In this work, we investigated whether a model that performs worse than the student can still provide beneficial supervision. Our findings support an affirmative answer that a weak teacher can serve as an informative contrastive reference alongside a stronger teacher. Building on this insight, cOPD selectively reinforces beneficial token preferences and suppresses harmful ones through dual-branch supervision. Experiments across three tasks and diverse model configurations support its effectiveness in improving student performance while suppressing undesirable response patterns. Further evidence suggests that weak teacher contrasts may encourage exploration of unique, valid responses.

We believe standard KL-based OPD does not explicitly verify whether teacher preferences improve task outcomes, so matching the teacher distribution may not yield the most beneficial training direction. This motivates selecting useful signals from raw teacher supervision. We view cOPD as a step toward this goal, using weak-teacher contrasts as complementary information to identify which token preferences to reinforce or suppress. Future work could extend cOPD to multiple teachers, combining guidance from strong teachers with complementary contrasts from diverse weak teachers to identify and suppress a broader range of failure modes.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing our proposed method. We did not use generative AI tools for research ideation. The authors manually reviewed the AI-assisted implementation for consistency with the intended method and checked its behavior in our experiments.

We also used generative AI tools to aid and polish writing. This included suggestions for preliminary drafts, alternative word choices, and sentence formulations to express our ideas more clearly, as well as grammar checks. We treated AI-suggested drafts as references and rewrote the manuscript sentence by sentence. The authors determined the final wording.

We reviewed all AI assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

We use publicly available Large Language Models and dataset for training and evaluation, and we have included all training hyperparameters in Appendix A. We will release our implementation code upon acceptance.

## REFERENCES

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

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc NM Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

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. *arXiv preprint arXiv:2607.05394*, 2026.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, et al. Rethinking on-policy distillation of large language models ii: One training example. *arXiv preprint arXiv:2609.04172*, 2026.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Zhengyang Tang, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations*, volume 2025, pp. 100540–100569, 2025.

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. *arXiv preprint arXiv:2608.19098*, 2026.

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.

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

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

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

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

Shizhen Li, Zhiyu Shen, Yuyin Lu, Yunhe Pang, Jielin Song, Yanghui Rao, and Fu Lee Wang. Rethinking reverse kl as adaptive entropy distillation. *arXiv preprint arXiv:2608.14685*, 2026a.

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*, 2026b.

Xinyu Liu, Kechen Jiao, Chunyang Xiao, Runsong Zhao, Junhao Ruan, Bei Li, Jiahao Liu, Qifan Wang, Xin Chen, Jingang Wang, et al. Teacher-guided policy optimization for on-policy reasoning distillation under large policy divergence. *arXiv preprint arXiv:2605.13230*, 2026.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Yifan Niu, Han Xiao, Dongyi Liu, Zelong Wang, Dihong Gong, Yasheng Wang, and Jia Li. Breaking the tokenizer barrier: On-policy distillation across model families. *arXiv preprint arXiv:2606.09456*, 2026.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. Mad-opd: Breaking the ceiling in on-policy distillation via multi-agent debate, 2026a. URL https://arxiv.org/abs/2605.01347.

Yifan Wang, Bolian Li, Junlin Wu, Zhaoxuan Tan, Zheli Liu, Ruqi Zhang, Ananth Grama, and Qingkai Zeng. Drift: Learning from abundant user dissatisfaction in real-world preference learning. In *International Conference on Learning Representations*, volume 2026, pp. 120457–120479, 2026b.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026c.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

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

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-strong on-policy distillation, 2026. URL https://arxiv.org/abs/2607.26246.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yige Yuan, Teng Xiao, Shuchang Tao, Xue Wang, Jinyang Gao, Bolin Ding, and Bingbing Xu. Incentivizing strong reasoning from weak supervision. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7138–7156, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026.

## A Implementation Details

Table 3 summarizes the training hyperparameters for mathematical reasoning, instruction following, and question answering with abstention. For evaluation, we use temperature 0.7 for Pass@16 and temperature 0.5 for Pass@1, with top-$p = 0.95$. For IF evaluation, we exclude prompts with invalid grader outputs for any compared model within each configuration, yielding common evaluation subsets of 6,900 prompts for Qwen3 and 6,906 for Llama from the original 7,000 held-out prompts.

Table 3: Training hyperparameters across the three tasks. Maximum response lengths are measured in tokens.

| Hyperparameter | Math (DAPO) | IF | SQuAD 2.0 |
|---|---|---|---|
| $\delta$ | 0.05 | 0.05 | 0.05 |
| $\tau$ | 1.0 | 1.0 | 1.0 |
| Optimizer | AdamW | AdamW | AdamW |
| Learning rate | $1 \times 10^{-6}$ | $1 \times 10^{-6}$ | $1 \times 10^{-6}$ |
| Weight decay | 0.01 | 0.01 | 0.01 |
| Training epochs | 3 | 3 | 3 |
| Prompt batch size | 64 | 64 | 64 |
| Responses per prompt | 4 | 4 | 4 |
| PPO mini-batch size | 64 | 64 | 64 |
| Training sampling temperature | 1.0 | 1.0 | 1.0 |
| Distillation support size | 16 | 16 | 16 |
| Maximum response length | 7,168 | 4,096 | 4,096 |

Table 4: Training data and supervision branch ablations on mathematical reasoning. The filtered subset contains 6,016 DAPO-Math-17k prompts, which is consistent with the main experiment. The positive only and negative only variants activate only $m_v^+$ and $m_v^-$, respectively. Bold and underlined scores indicate the best and second-best results within each training-data group.

| Method | AIME24 Pass@16 ↑ | AIME25 Pass@16 ↑ | AIME26 Pass@16 ↑ | AMC23 Pass@16 ↑ | Avg. Pass@16 ↑ | Omni-Math Pass@1 ↑ |
|---|---|---|---|---|---|---|
| *Full DAPO-Math-17k* | | | | | | |
| FKL | <u>76.67</u> | **60.00** | <u>36.67</u> | 89.16 | 72.83 | 18.54 |
| RKL | **80.00** | **60.00** | 33.33 | <u>93.98</u> | <u>75.14</u> | 19.32 |
| JSD | <u>76.67</u> | <u>53.33</u> | **43.33** | <u>93.98</u> | <u>75.14</u> | <u>19.53</u> |
| **cOPD (ours)** | **80.00** | **60.00** | <u>36.67</u> | **95.18** | **76.30** | **24.14** |
| *Filtered DAPO-Math-17k subset (6,016 prompts)* | | | | | | |
| FKL | 76.67 | 53.33 | <u>43.33</u> | 91.57 | 73.99 | 12.19 |
| RKL | 73.33 | 46.67 | 40.00 | <u>93.98</u> | 72.83 | 15.14 |
| JSD | 76.67 | <u>56.67</u> | <u>43.33</u> | 92.77 | 75.14 | 14.14 |
| cOPD ($m_v^+$ only) | **83.33** | **60.00** | <u>43.33</u> | 92.77 | <u>76.88</u> | 20.95 |
| cOPD ($m_v^-$ only) | 73.33 | 53.33 | 40.00 | 91.57 | 72.83 | **26.76** |
| **cOPD (ours)** | <u>80.00</u> | <u>56.67</u> | **46.67** | **95.18** | **77.46** | <u>23.82</u> |

## B Ablation Study

Table 4 examines training set selection and branch contributions of cOPD. When trained on the full DAPO-Math-17k dataset, cOPD exceeds the strongest single-teacher baselines in pooled Pass@16 (76.30% versus 75.14% for RKL and JSD) and Omni-Math Pass@1 (24.14% versus 19.53% for JSD). Training on the teacher-separated subset of 6,016 prompts yields 77.46% pooled Pass@16 and 23.82% Omni-Math Pass@1, exceeding every full-data single-teacher OPD baseline on both metrics despite using fewer student-training prompts. Full data cOPD remains competitive, indicating that its performance is not restricted to the filtered subset. Among the cOPD variants, positive only outperforms negative only on all four AIME/AMC Pass@16 benchmarks, whereas negative-only leads on Omni-Math Pass@1. Combining the branches gives the highest pooled Pass@16 among these variants and an intermediate Omni-Math score, balancing their different strengths across the two evaluation metrics.

## C Examples for instruction following

We show abbreviated responses to a shared prompt. The cOPD and RKL examples correspond to the highlighted cOPD-only cluster and circled word limit failure in Figure 2. Markdown markers are preserved literally, and `[...]` marks omitted text.

**Shared prompt**

Which avenue in France was renamed after Victor Hugo? Highlight at least 7 sections in your answer using Markdown, for example, `*highlighted section*`. Answer in fewer than 606 words. The letter v must appear at least 13 times in your response.

Table 5: cOPD response (all instruction constraints satisfied).

The avenue in France, `*Avenue Victor Hugo*`, was renamed after the famous writer Victor Hugo. `[...]` `*V*` is an important element in the avenue's name, reflecting the honor paid to him.

`*V*` also appears in the names of many other roads in Paris, such as `*Vétheuil*`, `*Vieux-Paris*`, and `*Ville-Victor*`. `[...]`

Table 6: RKL response (word-limit violation).

| |
|---|
| The avenue in France renamed after Victor Hugo is `**Avenue Victor Hugo**.` `[...]`<br><br>With Victor Hugo's presence, this avenue has become a symbol of French culture and history. `*highlighted section*` `[...]` With Victor Hugo's presence, this avenue has become a symbol of French culture and history. `*highlighted section*` `[...]` With the participation of many artists and writers, Avenue Victor Hugo has |

*Note.* The full recorded response contains extensive repetition and ends mid-sentence, as preserved above.

Table 7: MAD-OPD response.

| |
|---|
| The avenue in France renamed after Victor Hugo is `**Avenue Victor Hugo**.` `[...]`<br><br>`*highlighted section*`<br>Avenue Victor Hugo also passes through `**Rue de Rivoli Street**`, `**Rue de la Pompe Street**`, and `**Rue de la Pompe Street**.` `[...]` |

Table 8: W2S-OPD response.

| |
|---|
| The avenue in France renamed after Victor Hugo is `**Avenue Victor Hugo**.` `[...]`<br><br>`*highlighted section*`<br>The road also hosts many cultural and artistic events, such as art exhibitions and performances that take place regularly. `[...]` With 13 occurrences of the letter `**v**`, Avenue Victor Hugo is one of the most important roads in Paris, closely associated with French culture and history. |