COLLECTIVE WISDOM: ON-POLICY SELF-DISTILLATION WITH SYNERGISTIC MULTI-PERSPECTIVE TEACHER POLICY FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

In On-Policy Self-Distillation (OPSD), the student receives dense token-level supervisory signals from a teacher branch conditioned on privileged information, thereby improving its reasoning ability without relying on a stronger external teacher. However, a single teacher perspective can introduce bias and incomplete coverage in the supervisory signals, causing student updates to rely excessively on local supervisory signals and leading to performance instability or even collapse during the distillation process. Although incorporating multi-perspective teachers can broaden the sources of supervisory signals, conflicts among the supervisory signals provided by different teachers can make inadequate policy fusion susceptible to performance fluctuations during the distillation process. To address these challenges, we propose Collective Wisdom (CW), a synergistic multi-perspective teacher policy fusion method for OPSD. Specifically, we first take the union of the top-k vocabulary from each teacher to construct a joint policy space and normalize the individual teacher policies within this space to fuse them. We then realign the shared top-k vocabulary based on the fused policy to distill a consensus policy, which is employed to optimize the student policy. This design enables the synergistic fusion and stable distillation of multi-perspective teacher policies. Furthermore, experiments on models of different sizes and architectures across four mathematical reasoning benchmarks show that our method outperforms representative baselines in stabilizing distillation and mitigating performance collapse.

1 INTRODUCTION

With the continuous advancement of Large Language Models (LLMs) (Singh et al., 2025; Team, 2025; Qin et al., 2025; Zeng et al., 2026; Guo et al., 2025), they have demonstrated remarkable capabilities in semantic understanding, instruction following, and complex reasoning. However, LLMs exhibit limited performance when faced with complex reasoning tasks in specific domains. In recent years, many studies have shown that post-training can effectively improve the capabilities of LLMs on complex reasoning tasks in specific domains. Consequently, post-training has gradually become a core technique for improving the domain adaptability and reasoning capabilities of LLMs.

Post-training methods can be categorized into two types according to their optimization objectives: Policy Optimization (PO) and Knowledge Distillation (KD). PO methods (Schulman et al., 2017; Shao et al., 2024; Yu et al., 2026; Yue et al., 2025; Chen et al., 2025) optimize the model policy by constructing advantage estimates from rewards. KD methods (Gu et al., 2023; Agarwal et al., 2024; Kim & Rush, 2016; Yang et al., 2026b; Jin et al., 2026a; Ko et al., 2024; Wang et al., 2026b) transfer knowledge from the teacher model to the student model by aligning their distributions. Recently, many researchers have begun to explore OPSD methods (Zhao et al., 2026; Lu et al., 2026). In these methods, the student model samples responses on its own, while the same model equipped with privileged information serves as the teacher to provide dense token-level supervisory signals, thereby improving model performance without relying on a stronger external teacher. Many existing OPSD methods primarily adopt a single-teacher distillation strategy (Li et al., 2026a), while some studies (Wang et al., 2026a; Nguyen et al., 2026; Jin et al., 2026b) explore multiple teachers to broaden the sources of supervision. For example, MT-OPSD (Wang et al., 2026a) assigns equal weights to the distillation losses of multiple teachers.

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.

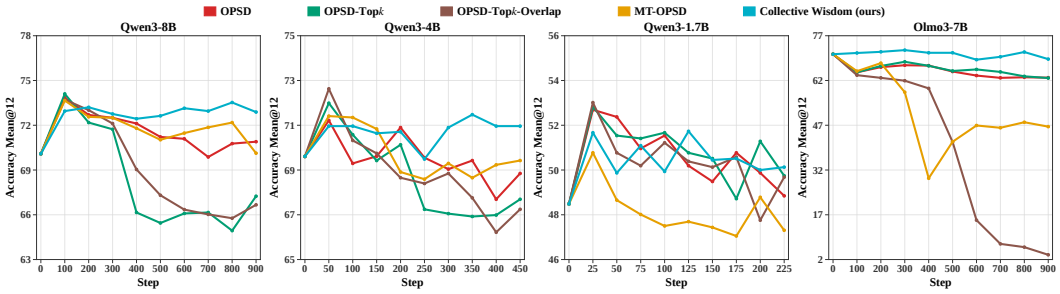


Figure 1: Overall Mean@12 accuracy across training steps for different methods using four backbone models on four mathematical benchmarks (AMC23, AIME24, AIME25, and HMMT25).

Despite these advances, stable distillation in OPSD still faces some challenges. Firstly, although a fixed privileged context can provide guidance for reasoning, it may steer the student toward a specific reasoning path. The resulting supervisory signals may be biased, thereby affecting the stability of distillation and even causing performance collapse. As shown in Figure 1, we observe that performance collapse occurs regardless of whether OPSD uses the full vocabulary, a top-k vocabulary, or the teacher-student intersection top-k vocabulary. This phenomenon is particularly pronounced when updates are restricted to only a small subset of the vocabulary. Secondly, although a multi-teacher configuration expands the sources of supervision, it also introduces a new policy-fusion problem: teachers conditioned on different privileged information yield conflicting policies, whose inadequate fusion destabilizes distillation. As shown in Figure 1, MT-OPSD exhibits a performance collapse during training, indicating that simply averaging teacher losses is insufficient to ensure training stability. Therefore, how to introduce teachers with multi-perspective privileged information and fuse their policies remains a key challenge in improving the distillation stability of OPSD.

To address these challenges, we propose Collective Wisdom, a synergistic multi-perspective teacher policy fusion method for on-policy self-distillation. Specifically, we construct three teacher perspectives conditioned on different privileged contexts based on responses self-sampled by the student, thereby obtaining rich supervisory signals. We then take the union of the top-k vocabulary from each teacher to construct a joint policy space from the policies of the individual teachers and fuse these teacher policies within this space to distill a multi-perspective consensus policy. Finally, we construct the distillation objective based on the consensus policy over the aligned top-k vocabulary to optimize the student policy, thereby achieving synergistic fusion and stable distillation of multi-perspective teacher policies. In addition, we conduct evaluations of four models of different sizes and architectures across four mathematical reasoning benchmarks. The experimental results show that our method outperforms representative baselines, reduces performance fluctuations during distillation while maintaining reasoning performance, and mitigates performance collapse.

Our main contributions are as follows:

- We propose a synergistic multi-perspective teacher policy fusion method that integrates teacher policies into a consensus policy within a joint policy space and distills it over an aligned top-k vocabulary, improving distillation stability.
- We construct three teacher perspectives conditioned on different privileged contexts based on responses self-sampled by the student, thereby obtaining rich supervisory signals.
- Experimental results on models with different sizes and architectures and across multiple mathematical reasoning benchmarks demonstrate that our method improves the distillation stability of OPSD and mitigates performance collapse.

2 Related Work

2.1 Knowledge Distillation

KD transfers the teacher's reasoning style and behavioral patterns to the student by aligning their conditional distributions along the distillation responses (Xu et al., 2024; Sanh et al., 2019; Lin

et al., 2020; Agarwal et al., 2024; Xu et al., 2026). Depending on whether the behavior policy matches the target policy, existing methods can be categorized into Off-Policy Distillation (OffPD) and On-Policy Distillation (OPD). OffPD samples responses offline and trains the student to fit the supervisory signals provided by the teacher on these responses, decoupling training from sampling (Guo et al., 2025; Gu et al., 2024; Mirzadeh et al., 2020; Yang et al., 2024). However, the distillation responses in these methods depend on the sampling model's behavior policy, concentrating supervisory signals on the state distribution induced by that policy. This distribution differs from the student's own state distribution during inference, preventing the student from receiving supervisory signals targeted at states where it is prone to errors (Agarwal et al., 2024). To address this problem, many studies have proposed OPD, which samples responses online from the student policy and trains the student using dense token-level supervisory signals provided by the teacher on these responses (Ko et al., 2024; Song & Zheng, 2026; Ye et al., 2025; Hou et al., 2026). For example, DistiLLM (Ko et al., 2024) reduces online sampling costs by adaptively reusing responses previously generated by the student. However, the benefits of OPD depend on compatibility between the teacher's and student's reasoning patterns, and the teacher must provide new capabilities that the student has not previously acquired; otherwise, distillation yields no gains and may even cause negative transfer (Li et al., 2026a; Fu et al., 2026). Moreover, employing a stronger teacher model incurs additional resource overhead (Zhao et al., 2026). Researchers have therefore proposed OPSD methods (Zhao et al., 2026; Yang et al., 2026a) to address these problems. These methods construct a teacher branch conditioned on additional privileged information to provide dense token-level supervisory signals to the student, thereby improving its reasoning ability without relying on a stronger external teacher.

2.2 On-Policy Self-Distillation

OPSD samples responses online from the student policy and constructs one or more teacher branch conditioned on additional privileged information to provide dense token-level supervisory signals, thereby improving the student's reasoning ability without relying on a stronger external teacher (Zhao et al., 2026; Shenfeld et al., 2026; Penaloza et al., 2026; Li et al., 2026b). Depending on the number of teachers used for distillation, existing methods can be categorized into single-teacher OPSD and multi-teacher OPSD. Single-teacher OPSD constructs a teacher branch conditioned on one form of privileged information to provide dense token-level supervisory signals to the student (Lu et al., 2026; Hübotter et al., 2026). However, a single teacher perspective can introduce bias and insufficient coverage in policy supervision, causing student updates to rely excessively on local policy supervision and leading to instability or even collapse during distillation (Nguyen et al., 2026). Multi-teacher OPSD constructs multiple teacher branches conditioned on different forms of privileged information to broaden supervisory coverage (Ma et al., 2026; Nguyen et al., 2026; Jin et al., 2026b). For example, AVSD (Nguyen et al., 2026) decomposes policy discrepancies between multi-perspective teachers and the student into a consensus vector and teacher-specific residual components, leveraging the consensus vector to control the update direction and the residual components to control the update magnitude. Although incorporating multi-perspective teachers can broaden the sources of policy supervision, these methods handle conflicting supervisory signals across perspectives through strategies such as filtering or scaling. Their inadequate policy fusion leaves conflicts among multi-perspective teacher policies unresolved, making training susceptible to substantial performance fluctuations. To address these challenges, we propose Collective Wisdom, a synergistic multi-perspective teacher policy fusion method for OPSD. Specifically, we construct a joint policy space for multiple teachers and fuse their policies to distill a consensus policy over the aligned candidate vocabulary for optimizing the student, thereby improving the stability of distillation.

3 Preliminaries

Multi-teacher OPSD (Nguyen et al., 2026) samples responses online from the student policy and constructs multiple teacher branches conditioned on privileged information to provide dense token-level supervisory signals, thereby improving the student's reasoning ability without relying on a stronger external teacher. Specifically, given a question x, the student model first autoregressively generates a response: $y = (y_1, \ldots, y_L) \sim p_\theta^S(\cdot \mid x)$, where L denotes the response length and θ denotes the student model parameters. The student and all teacher branches then compute conditional probabilities at the same prefix $y_{<l}$. The student conditions on the question x and the prefix $y_{<l}$,

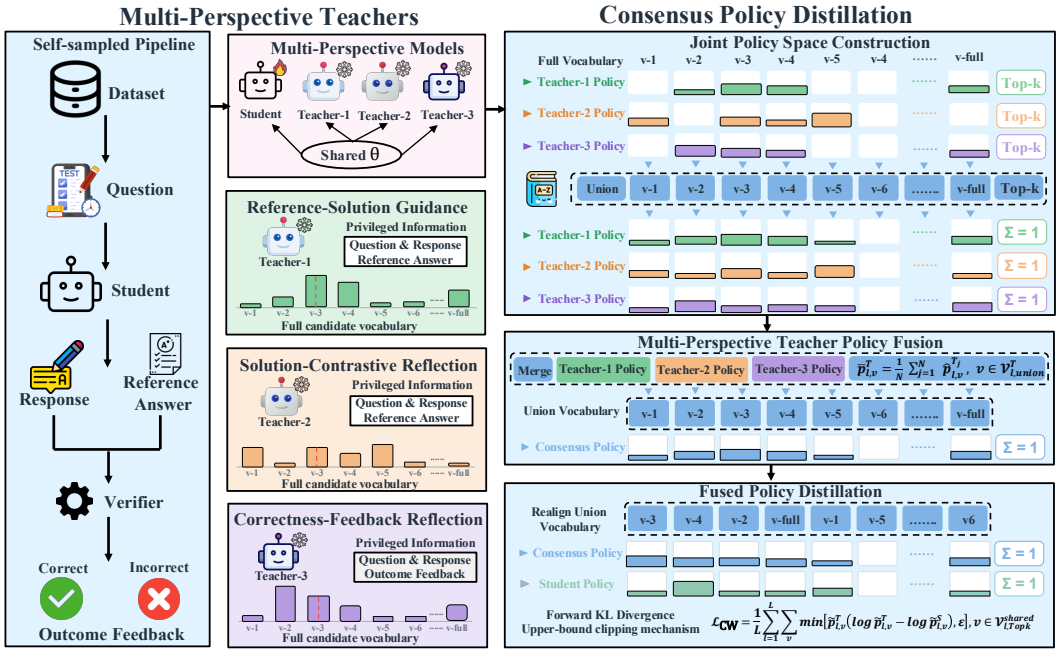


Figure 2: Overview of Collective Wisdom for synergistic multi-perspective teacher policy fusion in OPSD, which consists of two parts: multi-perspective teachers and consensus policy distillation.

while each teacher additionally conditions on its corresponding privileged information c_j:

$$p^S_{l,v} = p^S_\theta(v \mid x, y_{<l}), \quad p^{T_j}_{l,v} = p^{T_j}_\theta(v \mid x, c_j, y_{<l}), \tag{1}$$

where $l \in \{1, \ldots, L\}$ is the token position in the response, $j \in \{1, \ldots, N\}$ indexes the teacher branches, and $v \in \mathcal{V}$ is a token in the full vocabulary $\mathcal{V}$ shared by both the student and all teachers.

Finally, multi-teacher OPSD optimizes the student by minimizing the forward KL divergence from each teacher's conditional distribution to the corresponding student distribution, while applying an upper-bound ϵ clipping mechanism to prevent outlier terms from dominating training:

$$L_{\text{MT-OPSD}} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\sum_{v\in\mathcal{V}} \min\left[p^{T_j}_{l,v}\left(\log p^{T_j}_{l,v} - \log p^S_{l,v}\right), \epsilon\right] \tag{2}$$

During the optimization process, all teacher distributions are treated as fixed, and gradients are backpropagated only through the student probabilities, thereby internalizing the guidance provided by the different privileged contexts into a student policy that does not rely on this information.

4 Method

As shown in Figure 2, Collective Wisdom consists of two components: multi-perspective teachers leverage different privileged information to construct token-level supervisory signals based on responses self-sampled by the student; consensus policy distillation fuses the multi-perspective teacher policies into a consensus policy within a joint policy space and constructs a distillation objective over the aligned top-k candidate vocabulary to optimize the student policy.

4.1 Multi-Perspective Teachers

In OPSD, the teacher and student branches share the same model, differing only in their input information. This means that the privileged context influences both the additional information available to the teacher relative to the student and the learnable supervisory signals. Although a single privileged context can provide guidance, it steers the student toward a specific reasoning path. This

causes the student to rely excessively on local supervisory signals, leading to substantial fluctuations during distillation. This is particularly pronounced in top-k vocabulary distillation, because after top-k truncation the vocabulary concentrates on the high-probability region of a single teacher policy, making it difficult for the student to obtain supervision along other feasible directions.

To broaden the sources of supervisory signals, we construct different supervisory signals from three perspectives: reference-solution guidance, solution-contrastive reflection, and correctness-feedback reflection. Specifically, the student generates a response $y \sim p_\theta^S(\cdot \mid x)$ for a question x, and compares it against the reference answer r to obtain correctness feedback on the outcome. We then construct a set of teachers with different perspectives that share the backbone model, $\mathcal{T} = \{T_j\}_{j=1}^N$, where $N = 3$. The teacher T_1 from the reference-solution guidance perspective takes the reference answer and an instruction as its privileged context c_1, requiring the teacher to understand the reference answer before producing a reasoning solution. The teacher T_2 from the solution-contrastive reflection perspective takes the reference answer, the response, and a reflective instruction as its privileged context c_2, requiring the teacher to contrastively reflect on the differences between the reference answer and the response before producing a reasoning solution. The teacher T_3 from the correctness-feedback reflection perspective takes the outcome feedback, the student response, and a reflective instruction as its privileged context c_3, requiring the teacher to reflect on the reasoning process according to the outcome feedback, locate the problem, and refine the reasoning solution. Prompt templates for all three teachers are provided in Appendix A.1.

At the same position l in the student response, the student computes its policy conditioned on the original question x and the response prefix $y_{<l}$, whereas each teacher T_j additionally conditions on its corresponding privileged information c_j:

$$p^S_{l,v_s} = p^S_\theta(v_s \mid x, y_{<l}), \quad p^{T_j}_{l,v_j} = p^{T_j}_\theta(v_j \mid x, c_j, y_{<l}), \tag{3}$$

where v_s and v_j denote the vocabularies of the student and the individual teachers, respectively. Since different privileged contexts lead to inconsistent high-probability tokens, the probabilities assigned to the same token may differ substantially, resulting in conflicts among the different teacher policies and thereby affecting the stability of distillation. Therefore, how to introduce teachers with multi-perspective privileged information and effectively fuse multiple teacher policies to form reliable supervisory signals is a key problem in improving the distillation stability of OPSD.

4.2 Consensus Policy Distillation

To prevent distillation instability caused by conflicts among multi-perspective teacher policies, we first take the union of the top-k vocabularies of the teachers from all perspectives to construct a joint policy space and normalize the individual teacher policies within this space to fuse them. Then, we realign the shared top-k vocabulary according to the fused policy to distill the consensus policy, which is employed to optimize the student policy.

Joint Policy Space Construction First, at position l in the student response, we compute the union $\mathcal{V}^T_{l,\mathrm{Union}} = \bigcup_{j=1}^N \mathcal{V}^{T_j}_{l,\mathrm{Topk}}$ of the top-k candidate vocabularies $\mathcal{V}^{T_j}_{l,\mathrm{Topk}} = \mathrm{Topk}_{v_j \in \mathcal{V}_j}\left(p^{T_j}_{l,v_j}\right)$ of the individual teachers. This union vocabulary can retain the high-probability tokens of each perspective and provides a joint policy space for teacher policy fusion. Since different teachers assign different probabilities over the joint policy space, we perform local normalization on each of them separately:

$$\hat{p}^{T_j}_{l,v_j} = \frac{p^{T_j}_{l,v_j}}{\sum_{u \in \mathcal{V}^T_{l,\mathrm{Union}}} p^{T_j}_{l,u}}, \tag{4}$$

where $\hat{p}^{T_j}_{l,v_j}$ is the normalized policy of each teacher over the same joint policy space. Through this normalization, we represent the individual teacher policies within the same joint policy space, providing the basis for the subsequent policy fusion.

Multi-Perspective Teacher Policy Fusion Then, we fuse the normalized teacher policies within the joint policy space constructed above to obtain a multi-perspective consensus policy:

$$\bar{p}^T_{l,v} = \frac{1}{N} \sum_{j=1}^{N} \hat{p}^{T_j}_{l,v}, v \in \mathcal{V}^T_{l,\text{Union}}, \tag{5}$$

where $\bar{p}^T_{l,v}$ denotes the average degree of support that the teachers from all perspectives assign to token v. Although the previous joint policy space can guarantee that all teachers lie in the same joint policy space, the policy fusion changes the probability ordering of the consensus vocabulary. We realign the consensus vocabulary according to the fused policy $\mathcal{V}^{\text{shared}}_{l,\text{Topk}} = \text{Topk}_{v \in \mathcal{V}^T_{l,\text{Union}}} \left(\bar{p}^T_{l,v} \right)$ to ensure that it accurately reflects the consensus among these teachers.

Fused Policy Distillation Following the realignment of the consensus vocabulary, the fused policy and the student policy assign different probabilities within the new space, so we locally normalize them separately to construct a new joint policy space:

$$\tilde{p}^T_{l,v} = \frac{\bar{p}^T_{l,v}}{\sum_{u \in \mathcal{V}^{\text{shared}}_{l,\text{Topk}}} \bar{p}^T_{l,u}}, \quad \tilde{p}^S_{l,v} = \frac{p^S_{l,v}}{\sum_{u \in \mathcal{V}^{\text{shared}}_{l,\text{Topk}}} p^S_{l,u}}, \tag{6}$$

where $\tilde{p}^T_{l,v}$ and $\tilde{p}^S_{l,v}$ denote the consensus policy and the student policy over the vocabulary $\mathcal{V}^{\text{shared}}_{l,\text{Topk}}$.

Finally, we construct the distillation objective based on the forward KL divergence to optimize the student policy and adopt an upper-bound clipping mechanism to prevent excessively large local discrepancies between the teacher and the student:

$$\mathcal{L}_{\text{CW}} = \frac{1}{L} \sum_{l=1}^{L} \sum_{v \in \mathcal{V}^{\text{shared}}_{l,\text{Topk}}} \min \left[\tilde{p}^T_{l,v} \left(\log \tilde{p}^T_{l,v} - \log \tilde{p}^S_{l,v} \right), \epsilon \right]. \tag{7}$$

where ϵ is the clipping threshold. The student's vocabulary is aligned with the teacher's vocabulary.

In summary, we balance the divergence among the policies of different perspectives through policy fusion, ensuring that the supervisory signals during distillation focus on the region of the consensus policy, thereby improving the stability of distillation and mitigating performance collapse.

5 Experiments

5.1 Experimental Setup

Dataset Following OPSD (Zhao et al., 2026), we adopt OpenThoughts_math_30k as the training dataset, comprising approximately 30K problem–answer pairs drawn from the mathematical reasoning subset of OpenThoughts (Guha et al., 2026). We evaluate our method on four representative mathematical reasoning benchmarks: AMC23 (MAA, 2023), AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), and HMMT25 (Dekoninck et al., 2026).

Evaluation Metrics To assess the stability of the training process for all methods, we save nine model checkpoints at evenly spaced training steps. For each checkpoint, we generate 12 responses per question and compute the overall average accuracy (Mean@12). We then calculate the mean (Avg) and variance (Var) of these Mean@12 scores across the nine checkpoints as descriptive measures of average performance and variability during training.

Models We select four models spanning different sizes and architectures as our backbone models: Qwen3-8B, Qwen3-4B, Qwen3-1.7B (Team, 2025), and Olmo3-7B (Olmo et al., 2025).

Baselines We compare our method with nine representative baselines: (1) GRPO (Shao et al., 2024) computes advantages from the rewards of responses within the same group to update the policy; (2) OPSD (Zhao et al., 2026) constructs a teacher branch conditioned on privileged information

Table 1: Results of our method compared with nine representative baselines on four mathematical reasoning benchmarks. Avg and Var denote the mean and variance of the Mean@12 scores over nine uniformly spaced checkpoints of each baseline. These results demonstrate that our method effectively improves training stability and mitigates performance collapse.

Method	AMC23		AIME24		AIME25		HMMT25		Overall	
	Avg↑	Var↓	Avg↑	Var↓	Avg↑	Var↓	Avg↑	Var↓	Avg↑	Var↓
Qwen3-1.7B										
GRPO	84.35	2.01	46.79	3.88	**34.81**	2.66	22.22	2.79	49.91	1.00
OPSD	84.44	2.19	**49.35**	4.87	33.64	2.65	**24.32**	1.80	50.75	1.49
OPSD-Top$k$	**85.37**	2.22	48.21	2.04	34.60	3.12	24.10	1.71	**50.94**	1.26
OPSD-Top$k$-Overlap	85.05	0.76	47.56	7.39	33.64	3.85	23.86	1.21	50.41	1.70
MT-OPSD	80.74	1.75	44.63	7.36	33.36	2.01	22.93	1.17	48.13	1.18
MT-OPSD-Merge	83.36	3.07	47.69	**1.99**	33.64	2.18	22.84	1.98	49.69	0.71
MAD-OPSD	81.64	0.94	38.80	2.02	32.44	2.50	19.63	**1.08**	46.09	0.69
AVSD	84.19	2.98	46.39	4.12	34.07	**1.53**	23.55	5.23	49.91	0.69
RLCSD	76.02	26.85	36.98	46.96	27.62	27.49	17.04	9.65	42.23	24.93
Collective Wisdom	84.75	**0.75**	48.67	2.19	34.32	2.44	23.27	1.18	50.60	**0.47**
Qwen3-4B										
GRPO	96.27	**0.24**	73.12	2.02	63.61	2.25	41.11	2.45	70.66	0.44
OPSD	95.65	0.46	71.88	2.17	61.98	4.89	39.81	3.77	69.51	0.99
OPSD-Top$k$	94.49	2.82	71.11	6.96	60.77	5.43	39.69	3.27	68.67	3.20
OPSD-Top$k$-Overlap	95.05	1.73	70.77	6.01	61.08	6.68	39.85	5.30	68.87	3.13
MT-OPSD	95.25	0.79	72.75	**0.93**	62.04	8.02	40.43	**1.12**	69.74	1.15
MT-OPSD-Merge	96.20	1.88	71.70	2.44	62.75	1.82	40.96	1.59	70.08	0.43
MAD-OPSD	94.03	0.67	65.99	2.73	57.81	8.89	36.70	2.01	65.97	2.12
AVSD	96.27	0.76	73.18	1.42	**63.74**	2.52	**41.54**	4.55	**70.81**	0.49
RLCSD	95.97	0.41	68.18	6.57	53.77	19.12	38.67	9.96	66.60	5.30
Collective Wisdom	**96.57**	0.41	**73.36**	1.29	63.46	**0.93**	41.14	2.99	70.78	**0.26**
Qwen3-8B										
GRPO	95.32	0.73	**76.17**	1.75	66.82	2.87	43.09	2.56	72.27	0.79
OPSD	95.63	0.56	73.52	3.02	65.62	7.88	43.92	1.28	71.67	1.32
OPSD-Top$k$	91.06	15.01	70.77	13.32	62.19	20.86	41.27	7.03	68.23	10.57
OPSD-Top$k$-Overlap	92.96	4.39	71.79	11.66	62.56	22.82	40.22	8.93	68.89	9.18
MT-OPSD	95.30	0.80	75.00	2.42	66.79	1.39	42.75	4.59	71.91	0.89
MT-OPSD-Merge	94.98	0.66	76.02	3.19	67.28	1.54	44.44	3.12	72.55	0.71
MAD-OPSD	94.70	0.63	72.01	1.78	62.38	2.78	39.51	3.25	69.27	1.05
AVSD	**96.46**	0.46	75.03	2.41	67.31	3.76	43.33	2.95	72.53	0.45
RLCSD	93.80	1.28	69.75	6.52	55.40	37.55	38.21	18.26	66.56	10.12
Collective Wisdom	95.95	**0.08**	76.08	**1.05**	**67.53**	**0.86**	**44.54**	**1.13**	**72.94**	**0.09**
Olmo3-7B										
GRPO	95.69	1.42	70.00	**0.72**	64.35	11.34	41.27	8.04	69.97	1.77
OPSD	90.32	1.37	64.14	3.64	56.85	7.32	39.23	4.39	64.76	2.71
OPSD-Top$k$	90.67	0.91	64.35	3.57	57.90	7.36	40.34	**3.85**	65.42	2.69
OPSD-Top$k$-Overlap	53.63	1237.69	33.06	659.12	30.37	539.56	19.88	249.26	35.73	660.22
MT-OPSD	72.80	178.49	48.21	170.92	42.69	123.99	28.40	57.83	49.93	128.43
MT-OPSD-Merge	95.39	2.59	68.58	5.75	65.12	9.64	40.99	8.12	69.66	2.29
MAD-OPSD	96.37	3.38	65.58	6.25	62.13	6.14	37.99	9.20	67.89	2.53
AVSD	96.04	1.40	**71.61**	4.32	63.86	10.12	40.86	7.88	70.24	1.83
RLCSD	50.28	2069.38	35.62	1095.59	31.98	899.00	20.62	392.50	35.83	1089.18
Collective Wisdom	**96.92**	**0.32**	69.94	2.48	**65.65**	**1.99**	**41.98**	4.09	**70.80**	**1.16**

to distill the student over the full vocabulary; (3) OPSD-Top$k$ (Li et al., 2026a) constructs a teacher branch conditioned on privileged information to distill the student over the top-k vocabulary; (4) OPSD-Top$k$-Overlap (Li et al., 2026a) constructs a teacher branch conditioned on privileged information to distill the student over the intersection of the teacher's and student's top-k vocabularies; (5) MT-OPSD (Wang et al., 2026a) performs distillation by averaging the distillation losses of multiple privileged teachers with equal weights; (6) MT-OPSD-Merge performs distillation using the con-

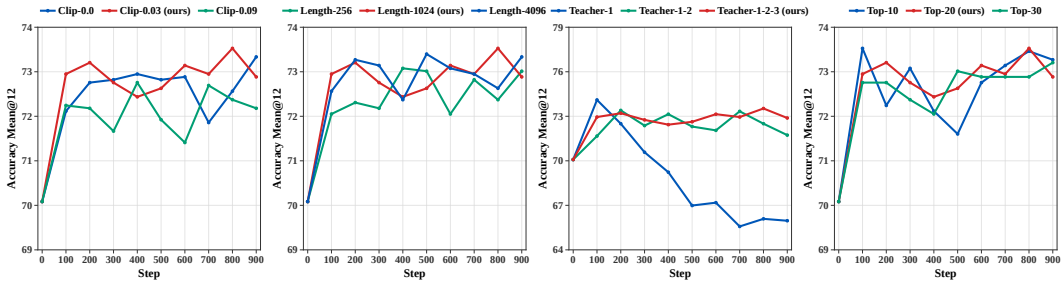


Figure 3: Overall Mean@12 accuracy across training steps for different hyperparameter settings using Qwen3-8B on four mathematical benchmarks (AMC23, AIME24, AIME25, and HMMT25).

sensus policy of multiple privileged teachers without realigning the consensus vocabulary, relying instead on the top-20 tokens selected from the deduplicated union of their original candidate vocabularies. (7) MAD-OPSD (Wang et al., 2026a) constructs privileged information through self-debate among multiple teachers for distillation; (8) AVSD (Nguyen et al., 2026) controls the direction and strength of distillation through multi-perspective consensus and discrepancy measures; (9) RLCSD (Pan et al., 2026) contrasts correct and incorrect privileged hints to modulate the direction and magnitude of updates. Detailed training and evaluation configurations are provided in Appendix A.2.

5.2 Main Results

As shown in Table 1, CW achieves the highest Overall Avg of 72.94 on Qwen3-8B. This represents gains of 3.67, 6.38, and 0.41 percentage points over MAD-OPSD, RLCSD, and AVSD, respectively. CW also reduces Overall Var from 1.05, 10.12, and 0.45 to 0.09. Moreover, CW achieves the lowest variance on all four benchmarks, demonstrating simultaneous improvements in average performance and training stability. On Qwen3-4B, CW achieves an Overall Avg of 70.78, exceeding OPSD and GRPO by 1.27 and 0.12 percentage points, respectively. It also reduces Overall Var from 0.99 and 0.44 to 0.26. Compared with OPSD, CW improves both average performance and stability. Compared with GRPO, it further reduces training fluctuations while maintaining similar average performance. In summary, CW achieves the lowest Overall Var across all four models while maintaining competitive average performance. These experimental results demonstrate that our method can effectively stabilize distillation and mitigate performance collapse.

5.3 Ablation Studies

To validate the effectiveness of each component in Collective Wisdom, we conduct a progressive ablation study against four baselines: (1) OPSD (w/o top-k vocabulary distillation); (2) OPSD-Top$k$ (w/o multi-teacher distillation); (3) MT-OPSD (w/o multi-teacher policy fusion); (4) MT-OPSD-Merge (w/o consensus vocabulary realignment). As shown in Table 1, CW improves both average performance and training stability compared with single-teacher distillation methods. Specifically, on Qwen3-8B, CW improves Overall Avg by 1.27 and 4.71 percentage points over OPSD and OPSD-Top$k$, respectively, while reducing Overall Var from 1.32 and 10.57 to 0.09. These results demonstrate its performance and stability advantages over single-teacher distillation. Furthermore, on Qwen3-4B, CW improves Overall Avg by 1.04 and 0.70 percentage points over MT-OPSD and MT-OPSD-Merge, respectively, while reducing Overall Var from 1.15 and 0.43 to 0.26. These results further validate the effectiveness of multi-teacher policy fusion and consensus vocabulary realignment. In summary, CW achieves the lowest Overall Var across all four models, with Overall Avg exceeding or approaching that of the other ablation variants. The ablation results indicate that CW improves distillation stability while maintaining competitive average performance, potentially by reducing supervision noise from individual teachers through multi-teacher consensus.

5.4 Hyperparameter Analysis

To ensure a fair comparison, we vary only a single hyperparameter at a time while keeping other parameters consistent with the main experiments. We conduct four sets of hyperparameter exper-

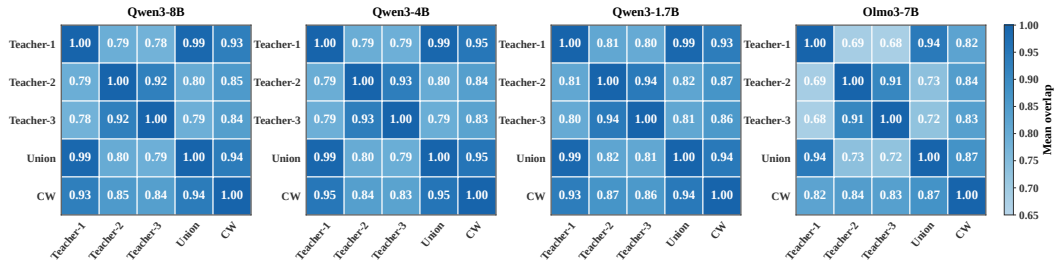


Figure 4: Heatmaps of vocabulary overlap among different teacher policies across four models.

iments to examine the clipping threshold (0.0, 0.03, and 0.09), response length (256, 1024, and 4096), the top-k set (10, 20, and 30), and the number of teachers (1, 2, and 3). As shown in Figure 3, performance fluctuations are more pronounced with Clip-0.0, while Clip-0.09 yields lower overall performance. Clip-0.03 achieves both high performance and training stability. A response length of 256 yields relatively low performance, whereas 1024 and 4096 achieve comparable performance. However, training with a response length of 4096 takes 3.78 times as long as training with 1024. As the number of teachers increases, the performance degradation observed in the later stages of single-teacher training is substantially mitigated. Increasing the number of teachers from two to three further stabilizes performance, suggesting the effectiveness of the different teacher perspectives. Furthermore, Top-10 uses a smaller candidate vocabulary for updates and exhibits more pronounced performance fluctuations. Top-30 yields slightly lower overall performance than Top-20. Overall, these results demonstrate the hyperparameter settings adopted in our main experiments.

5.5 Vocabulary Analysis

To analyze the impact of multi-perspective teacher policy fusion on the top-k candidate vocabulary, we record distillation information at all token positions in one response every 25 steps during a single training run. From these records, we extract the candidate vocabularies of the three teachers, MT-OPSD-Merge (Union), and Collective Wisdom (CW). At each position, we compute the overlap rate between each pair of candidate vocabularies. As shown in Figure 4, Teacher-1 has vocabulary overlap rates of 79% and 78% with Teacher-2 and Teacher-3 on Qwen3-8B, respectively. After policy fusion and vocabulary realignment, CW maintains a 93% overlap with Teacher-1, while its overlap rates with the other two teachers reach 85% and 84%, respectively. This indicates that CW retains most of Teacher-1's candidate tokens while incorporating complementary candidates from the other teachers. Furthermore, Union without vocabulary realignment has overlap rates of 99%, 80%, and 79% with the three teachers on Qwen3-8B, respectively. Its high overlap with Teacher-1 suggests limited incorporation of complementary candidates from the other perspectives. These results further validate the effectiveness of vocabulary realignment in incorporating complementary candidates across teacher perspectives. Overall, all four models exhibit a consistent trend, indicating that multi-perspective policy fusion preserves shared candidates while incorporating candidate tokens not covered by a single perspective. Under a fixed top-k candidate budget, this provides more consistent supervisory signals to the student, offering a vocabulary-level explanation for improved distillation stability and reduced performance collapse. The case study is in Appendix A.3.

6 Conclusion

In this paper, we propose Collective Wisdom, a synergistic multi-perspective teacher policy fusion method for OPSD. Specifically, we first take the union of the top-k candidate vocabulary from all teacher perspectives to construct a joint policy space and normalize the individual teacher policies within this joint policy space to fuse them. We then realign the shared top-k candidate vocabulary based on the fused policy to distill a multi-teacher consensus policy, which is employed to optimize the student policy. This design enables the synergistic fusion and stable distillation of multi-perspective teacher policies. Extensive experimental results on models with different sizes and architectures and across multiple mathematical reasoning benchmarks demonstrate that our method improves the distillation stability and mitigates performance collapse during training.

AI USE STATEMENT

In this work, we used generative AI tools for translation and language editing to improve the clarity and readability of the paper. Additionally, we used generative AI tools to help modify the flowchart of our method to make the proposed method easier to understand. We have reviewed and revised all AI-assisted work to ensure that it accurately reflects our research methods, experimental results, and conclusions. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support the reproducibility of this work, we provide detailed descriptions of the proposed method and experimental setup in the main text and appendix. The implementation code is included in the supplementary materials to enable readers to reproduce our experimental results.

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.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. *arXiv preprint arXiv:2605.00674*, 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.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Miniplm: Knowledge distillation for pre-training language models. *arXiv preprint arXiv:2410.17215*, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

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*, 2026a.

Yiqiao Jin, Yiyang Wang, Lucheng Fu, Yijia Xiao, Yinyi Luo, Haoxin Liu, B Aditya Prakash, Josiah Hester, Jindong Wang, and Srijan Kumar. Unisd: Towards a unified self-distillation framework for large language models. *arXiv preprint arXiv:2605.06597*, 2026b.

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.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

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.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision. *arXiv preprint arXiv:2608.06296*, 2026b.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, 2020.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, et al. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 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.

MAA. AMC 2023, 2023. URL `https://huggingface.co/datasets/math-ai/amc23`. Dataset compiled from the 2023 American Mathematics Competitions.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Duy Nguyen, Hanqi Xiao, Archiki Prasad, Zaid Khan, Anirban Das, Austin Zhang, Sambit Sahu, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Avsd: Adaptive-view self-distillation by balancing consensus and teacher-specific privileged signals. *arXiv preprint arXiv:2605.20643*, 2026.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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

Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

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. *arXiv preprint arXiv:2605.01347*, 2026a.

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*, 2026b.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

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.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

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*, 2026b.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1028–1043, 2024.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

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, 2026.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

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.

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.

A APPENDIX

A.1 PROMPT TEMPLATES FOR MULTI-PERSPECTIVE TEACHERS

The Teacher Prompt Template for the Reference-Solution Guidance Perspective

```
Problem: {problem}

Here is a reference solution to this problem:
=== Reference Solution Begin ===
{reference_solution}
=== Reference Solution End ===

After reading the reference solution above, make sure you truly understand
the reasoning behind each step — do not copy or paraphrase it. Now, using
your own words and independent reasoning, derive the same final answer to
the problem above. Think step by step, explore different approaches, and
don't be afraid to backtrack or reconsider if something doesn't work out:

Please reason step by step, and put your final answer within \boxed{}.
```

Figure 5: Prompt template for the reference-solution guidance teacher.

As shown in Figure 5, the teacher T_1 from the reference-solution guidance perspective takes the reference answer and an instruction as its privileged context, requiring the teacher to understand the reference answer before producing a reasoning solution.

The Teacher Prompt Template for the Solution-Contrastive Reflection Perspective

```
Problem: {problem}

Here is the student's attempted solution:
=== Student Solution Begin ===
{student_response}
=== Student Solution End ===

Here is a reference solution to this problem:
=== Reference Solution Begin ===
{reference_solution}
=== Reference Solution End ===

After reading the student's attempted solution and the reference solution
above, carefully compare the two. Identify which parts of the student's
reasoning are valid, which parts are incomplete or incorrect, and what
reasoning should be corrected. Do not copy the reference solution verbatim.
Do not simply repeat the student's answer. Now, using your own independent
reasoning, solve the original problem again. If the student's answer is
wrong, correct the reasoning and derive the correct final answer. If the
student's answer is right, provide a clean independent derivation that
justifies it. Think step by step, explore the reasoning carefully, and
backtrack if necessary:

Please reason step by step, and put your final answer within \boxed{}.
```

Figure 6: Prompt template for the solution-contrastive reflection teacher.

As shown in Figure 6, the teacher T_2 from the solution-contrastive reflection perspective takes the reference answer, the response, and a reflective instruction as its privileged context, requiring the teacher to contrastively reflect on the differences between the reference answer and the response before producing a reasoning solution.

The Teacher Prompt Template for the Correctness-Feedback Reflection Perspective

```
Problem: {problem}

Here is the student's attempted solution:
=== Student Solution Begin ===
{student_response}
=== Student Solution End ===

Evaluation result of the student's final answer:
=== Evaluation Result Begin ===
{eval_result}
=== Evaluation Result End ===

The evaluation result above tells whether the student's final answer is
correct. Use this feedback to reflect on the student's attempted solution.
If the result says the answer is wrong, carefully inspect the student's
reasoning, locate the likely mistakes, and solve the problem independently
to obtain the correct answer. If the result says the answer is right,
verify the student's reasoning and provide a clear independent derivation
of the same final answer. Do not simply repeat the student's response. Do
not rely only on the correctness label; reason through the problem
yourself. Think step by step, revise the reasoning if necessary, and derive
the final answer:

Please reason step by step, and put your final answer within \boxed{}.
```

Figure 7: Prompt template for the correctness-feedback reflection teacher.

As shown in Figure 7, the teacher T_3 from the correctness-feedback reflection perspective takes the outcome feedback, the student response, and a reflective instruction as its privileged context, requiring the teacher to reflect on the reasoning process according to the feedback, locate the problem, and refine the reasoning solution.

A.2 Experimental Setup

Common experimental hyperparameters Table 2 summarizes the common training and evaluation hyperparameters, with batch sizes varying across models.

Table 2: Common training and evaluation hyperparameters.

Hyperparameter	Value
Backbone models	Qwen3-8B, Qwen3-4B, Qwen3-1.7B, Olmo3-7B
Reasoning Mode	Thinking
GPU	NVIDIA 8 $*$ H800
Batch size	Qwen3-8B: 16; Qwen3-4B: 32; Qwen3-1.7B: 64; Olmo3-7B: 16
Gradient accumulation steps	2
Total batch size	Qwen3-8B: 32; Qwen3-4B: 64; Qwen3-1.7B: 128; Olmo3-7B: 32
Optimizer	AdamW
Learning rate	5×10^{-6}
Training epochs	1
Training maximum gradient norm	0.1
Training maximum completion length	1,024
Training maximum length	12,000
Training thinking mode	Disabled
VLLM GPU memory utilization	0.6
Precision	BF16
LoRA rank / alpha	64 / 128
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Checkpoint saving interval	Qwen3-8B: 100; Qwen3-4B: 50; Qwen3-1.7B: 25; Olmo3-7B: 100
Sampling temperature	1.0
Sampling top-p / top-k	0.95 / 20
Evaluation temperature	1.0
Evaluation top-p / top-k	0.95 / -1
Evaluation thinking mode	Enabled
Samples per query	12
Evaluation maximum completion length	38,912
Evaluation Maximum model context length	40,960
Evaluation GPU memory utilization	0.9

Experimental hyperparameters for different methods Table 3 lists the hyperparameters for different methods used in our experiments, including all nine baselines and Collective Wisdom.

Table 3: Experimental hyperparameters for different methods.

Method	Hyperparameter	Value
GRPO	Rollouts per prompt	8
	KL regularization coefficient	0.001
	Policy clipping threshold	0.2
OPSD	Rollouts per prompt	1
	Distillation vocabulary	Full vocabulary
	Divergence	Forward KL
	KL clipping	0.03

Table 3: Experimental hyperparameters for different methods.

Method	Hyperparameter	Value
OPSD-Top$k$	Rollouts per prompt	1
	Distillation vocabulary	Teacher top-20
	Divergence	Forward KL
	KL clipping	0.03
OPSD-Top$k$-Overlap	Rollouts per prompt	1
	Distillation vocabulary	Student-teacher top-20 intersection
	Divergence	Forward KL
	KL clipping	0.03
MT-OPSD	Rollouts per prompt	1
	Number of teachers	3
	Divergence	Forward KL
	KL clipping	0.03
MT-OPSD-Merge	Rollouts per prompt	1
	Number of teachers	3
	Divergence	Forward KL
	KL clipping	0.03
MAD-OPSD	Rollouts per prompt	1
	Teacher configuration	Two teachers sharing the student model
	Debate rounds	2
	Dynamic teacher weighting	Enabled
	Initial teacher weights	0.5, 0.5
	Distillation vocabulary	Teacher top-20
	JSD mixing coefficient	0.5
	On-policy coefficient	1.0
	Sequence-level KD	Disabled
	Student repetition penalty	1.2
AVSD	Rollouts per prompt	1
	Number of teachers	3
	Privileged information	Full/partial solution, final answer
	Partial-solution retention ratio	0.5
	Privileged information weighting	Uniform
	Distillation vocabulary	Per-teacher top-20
	JSD interpolation coefficient	0.5
	JSD clipping threshold	0.05
	Repetition penalty	1.2
	Exponential moving average teacher	Disabled
	Epistemic-preservation regularizer	Disabled
RLCSD	Rollouts per prompt	8
	Maximum negative hints	4
	Contrastive-signal temperature	0.02
	Modulation scale	0.5
	Token-selection threshold	0.02
	Modulated-path weight	1.0
	Policy clipping threshold	0.2
	KL regularization coefficient	0.001
	Residual scaling coefficient	1.0
	Residual clipping range	$[-2.0, 2.0]$
	Importance-sampling threshold	2.0

Table 3: Experimental hyperparameters for different methods.

Method	Hyperparameter	Value
Collective Wisdom	Rollouts per prompt	1
	Number of teachers	3
	Teacher configuration	Three teachers sharing the student model
	Exponential moving average teacher	Disabled
	Distillation vocabulary	Multi-teacher consensus top-20
	Divergence	Forward KL
	KL clipping	0.03

Training time costs Table 4 summarizes training time costs, batch sizes, checkpoint intervals, and total training steps across different methods and different backbone models.

Table 4: Training time costs across different methods and models.

Method	Model	Training time	Batch size	Checkpoint interval	Total steps
GRPO	Qwen3-1.7B	0.6 h	128	25	228
	Qwen3-4B	1.9 h	64	50	456
	Qwen3-8B	5.3 h	32	100	911
	Olmo3-7B	6.0 h	32	100	911
OPSD	Qwen3-1.7B	0.6 h	128	25	228
	Qwen3-4B	1.7 h	64	50	456
	Qwen3-8B	4.6 h	32	100	911
	Olmo3-7B	4.5 h	32	100	911
OPSD-Top$k$	Qwen3-1.7B	0.5 h	128	25	228
	Qwen3-4B	1.7 h	64	50	456
	Qwen3-8B	4.7 h	32	100	911
	Olmo3-7B	4.5 h	32	100	911
OPSD-Top$k$-Overlap	Qwen3-1.7B	0.5 h	128	25	228
	Qwen3-4B	1.7 h	64	50	456
	Qwen3-8B	4.7 h	32	100	911
	Olmo3-7B	4.5 h	32	100	911
MT-OPSD	Qwen3-1.7B	0.6 h	128	25	228
	Qwen3-4B	2.0 h	64	50	456
	Qwen3-8B	5.0 h	32	100	911
	Olmo3-7B	4.8 h	32	100	911
MT-OPSD-Merge	Qwen3-1.7B	1.2 h	128	25	228
	Qwen3-4B	2.5 h	64	50	456
	Qwen3-8B	5.5 h	32	100	911
	Olmo3-7B	5.4 h	32	100	911
MAD-OPSD	Qwen3-1.7B	21.2 h	128	25	228
	Qwen3-4B	85.0 h	64	50	456
	Qwen3-8B	112.0 h	32	100	911
	Olmo3-7B	119.2 h	32	100	911
AVSD	Qwen3-1.7B	1.4 h	128	25	228
	Qwen3-4B	5.8 h	64	50	456
	Qwen3-8B	8.5 h	32	100	911
	Olmo3-7B	8.1 h	32	100	911
RLCSD	Qwen3-1.7B	2.5 h	128	25	228
	Qwen3-4B	5.5 h	64	50	456
	Qwen3-8B	14.3 h	32	100	911
	Olmo3-7B	13.8 h	32	100	911

Table 4: Training time costs across different methods and models.

Method	Model	Training time	Batch size	Checkpoint interval	Total steps
Collective Wisdom	Qwen3-1.7B	0.9 h	128	25	228
	Qwen3-4B	2.3 h	64	50	456
	Qwen3-8B	5.5 h	32	100	911
	Olmo3-7B	5.3 h	32	100	911

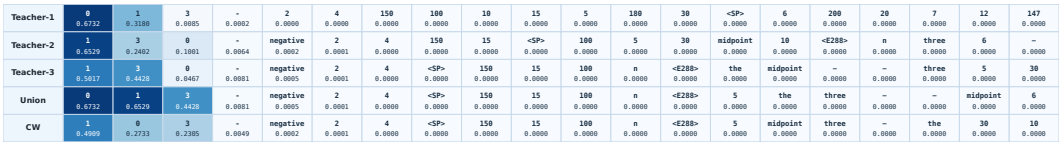


Figure 8: Heatmap of the top-20 tokens from each teacher at the same position in the response.

A.3 Case Study

Figures 8, 9, and 10 illustrate how Collective Wisdom (CW) fuses multi-perspective teacher policies into a shared distillation target. Immediately after `What about n=`, Teacher-1 lists `0` first, whereas Teacher-2 and Teacher-3 list `1` first. The token `1` ranks second, first, and first across the three teachers, with probabilities of 0.3180, 0.6529, and 0.5017, respectively. Union scores each token by its maximum probability across teachers, ranking `0` first and `1` second. Although Teacher-1 assigns `0` a probability of 0.6732, Teacher-2 and Teacher-3 assign it only 0.1001 and 0.0467. CW instead promotes `1` to first with probability 0.4909 and moves `0` to second with probability 0.2733. This reranking changes the supervision distribution to favor the student's sampled `1`, while retaining `0` as a plausible alternative. The favored numeral has a concrete role in verifying an inequality boundary. The problem asks for the sum of all integer solutions to $|n| < |n-3| < 9$. The recorded derivation solves $|n| < |n-3|$ over the reals, obtaining $n < 3/2$, and checks $n = 1$ and $n = 2$. With the recorded suffix `.5` held fixed, `1` gives $n = 1.5$, where $|1.5| = |1.5-3|$. This check explicitly verifies that the boundary is excluded by the strict inequality. Substituting `0` before the same suffix instead gives $n = 0.5$, where $0.5 < 2.5$, testing another interior point. CW therefore assigns greater relative supervision weight to the candidate corresponding to boundary verification. This comparison is an analytical substitution under a fixed suffix; it does not establish an improvement in generated solutions or training stability. Overall, the example illustrates how fusion can prioritize a token supported by all three teachers at a mathematically relevant verification step, even when Union ranks another candidate first.

Case Study

Query

What is the sum of all integer solutions to $|n| < |n-3| < 9$?

Response

<think>
I need to find the sum of all integer solutions to the inequality |n| < |n - 3| < 9. First, I need to understand what this inequality means. It says that the absolute value of n is less than the absolute value of (n minus 3), and that this is also less than 9. Since it's a chain inequality, I should break it down into two separate inequalities.

So, |n| < |n - 3| and |n - 3| < 9. But actually, since both have to hold, and |n - 3| is common, I need to satisfy both conditions simultaneously. Specifically, |n| < |n - 3| must be true, and |n - 3| < 9 must also be true.

I recall that |a| < |b| can be interpreted as the point a is closer to zero than b is. In this case, |n| < |n - 3| means that n is closer to zero than n-3 is. That might depend on whether n is positive or negative.

I could solve the inequality |n| < |n - 3| first, and then solve |n - 3| < 9, and find the intersection of the solutions.

Let me start with |n| < |n - 3|. This inequality represents the set of points where n is closer to 0 than to 3. In one dimension, this should be the midpoint or something.

I know that |a| < |b| implies that a^2 < b^2, if I square both sides, but only if both sides are non-negative, which they are since absolute values. Actually, for inequalities, |a| < |b| is equivalent to a^2 < b^2, because squaring is monotonic for absolute values.

Let me confirm that. If |a| < |b|, then squaring both sides, since both sides are non-negative, we get a^2 < b^2. Conversely, if a^2 < b^2, since squares are non-negative, taking square roots, we have |a| < |b|. Yes, so that's true.

So, |n| < |n - 3| implies n^2 < (n - 3)^2.

Let me expand that: n^2 < n^2 - 6n + 9.

Subtract n^2 from both sides: 0 < -6n + 9.

So, 6n < 9, which means n < 9/6, or n < 3/2.

Is that right? 0 < -6n + 9, so adding 6n to both sides: 6n < 9, so n < 3/2. Yes.

But is that sufficient? When I squared, I assumed that the inequality holds without considering the signs, but in this case, since absolute values are involved, and I derived it algebraically, it should be fine. But let me verify with a point.

Suppose n = 1, which is less than 1.5. |1| = 1, |1-3| = | -2 | = 2, and 1 < 2, good.

n = 2, |2| = 2, |2-3| = | -1 | = 1, but 2 < 1? No, 2 is not less than 1. But n=2 is less than 1.5? 2 > 1.5, so according to my inequality, it shouldn't satisfy, and indeed it doesn't.

Figure 9: A case study of Collective Wisdom, including the query and the response.

What about n=1.5? |1.5| = 1.5, |1.5-3| = | -1.5 | = 1.5, so 1.5 < 1.5 is false, so it shouldn't be included. Good.

But in this derivation, I got n < 3/2. But what about n negative? Let me try n=0. |0|=0, |0-3|=3, 0<3, true. n=0 < 1.5, good.

n=-1. | -1 | =1, | -1 -3 | = | -4 | =4, 1<4, true. n=-1 < 1.5, good.

n=3. |3|=3, |3-3|=0, 3<0? False, and n=3>1.5, not included. But what about n larger? n=4, |4|=4, |4-3|=1, 4<1? False. But according to n<1.5, it's excluded, correct.

Is there any n where |n| < |n

Figure 10: A case study of Collective Wisdom, including the query and the response.