# EVERY TOKEN NEEDS ITS OWN TEACHER: MULTI-TEACHER ADAPTIVE ON-POLICY SELF-DISTILLATION

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

## ABSTRACT

On-policy Self-Distillation (OPSD) has proven effective for post-training large language models by constructing a teacher conditioned on privileged references. However, existing approaches keep the teacher fixed either throughout training by consistently exposing it to the full reference, or within each response by adaptively varying partial reference exposure only across responses. In this paper, we argue that neither of them is sufficient. Instead, every token needs its own teacher, as the teacher preference can vary token by token within the same response. Through a teacher-intervention pilot study, we validate that the teacher preference differs across tokens, and reveal that a preferred teacher can be identified for each token through majority voting and KL-based selection. Motivated by these findings, we propose Multi-teacher Adaptive On-policy Self-Distillation (MTA-OPSD), the first token-level adaptive OPSD method. Our method constructs a multi-teacher pool consisting of candidate teachers, and enables every token to adaptively select its preferred teacher during training. Experiments with Qwen3 models on mathematical reasoning tasks show that MTA-OPSD consistently outperforms existing OPSD approaches with an average gain of up to $+5.9$ over the base model, highlighting the importance of adapting teacher supervision at the token level.

## 1 INTRODUCTION

On-policy Self-distillation (OPSD) (Zhao et al., 2026; Hübotter et al., 2026; Shenfeld et al., 2026; Cui et al., 2026) has emerged as a promising paradigm for post-training large language models (LLMs), showing superiority over other post-training paradigms like Supervised Fine-tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR) (e.g., GRPO (Guo et al., 2025)). In OPSD, the same model serves as both student and teacher. The student generates rollout responses, while the teacher provides supervision with access to privileged references, such as reference solutions. Compared with RLVR, OPSD provides dense token-level learning signals beyond sparse outcome rewards. Compared with SFT, it trains on trajectories sampled by the student itself, mitigating the exposure bias (Bengio et al., 2015).

Despite these advantages, recent studies (Kim et al., 2026; Kaur et al., 2026; Harne et al., 2026; Armandpour et al., 2026) reveal that OPSD can also suffer from degraded reasoning performance, which stems from the highly informative references exposed to the teacher. It may induce overly sharp teacher distributions, producing supervision that is poorly aligned with the student's learning dynamics (e.g., suppressing uncertainty (Kim et al., 2026; Zhang et al., 2026)). A line of work (Tan & Hong, 2026; Han et al., 2026b; Yang et al., 2026c) has attempted to address this student-teacher mismatch through adaptive reference exposure. Rather than providing the teacher with the complete reference, it adjusts how much of the reference the teacher receives based on the student's learning dynamics, such as the student-teacher gap. However, these methods adapt at the response level. Although reference exposure may vary across responses, it remains fixed across tokens within each response. In other words, all tokens in a rollout are supervised by the same teacher.

In this paper, we challenge this consensus by arguing:

> *The adaptive teacher should operate at the token level, because every token requires its own dedicated teacher.*



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.

A teacher that provides useful guidance at one token may be less suitable at another, even within the same response. Rather than treating it with the same teacher, each token in the response merits consideration with its own preferred teacher. As illustrated in Figure 1, when different teachers are applied to the same response, they can provide similar learning signals for some tokens but disagree on the direction of supervision for others. This suggests that not only each response, but also each token, may have distinct teacher preferences.

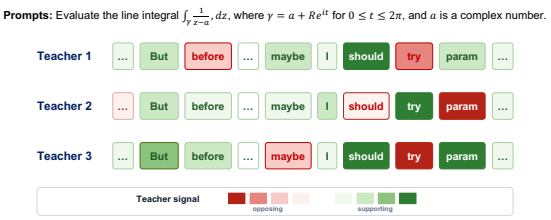


Figure 1: Token-level learning signals induced by different teachers for the same response. Teachers agree at some positions but provide opposing signals at others, motivating token-level teacher selection. Additional details on this visualization are provided in Appendix A.

While some prior work(Wang et al., 2026c; Hao et al., 2026; Xu et al., 2026b; Ke et al., 2026; Wei et al., 2026; Wang et al., 2026a; Li et al., 2026; Han et al., 2026a) has explored token selection for post-training (e.g., focusing solely on high-entropy tokens (Wang et al., 2026c)), how to validate our claim is challenging, and how to adaptively identify the preferred teacher for each token remains unexplored. To answer these, we conduct a teacher-intervention pilot study (Section 4) to quantify the token-level teacher preference. As shown in Figure 2, we preserve a student-generated prefix, allow a candidate teacher to generate a short continuation at a selected token position, and then return control to the student to complete the response. The accuracy gain from this "student-teacher-student" relay over pure student generation serves as a proxy for the teacher preference at that token position. A large gain implies a compatible teacher for that position, while a small gain suggests the opposite. Interestingly, the results are simple, yet clear and compelling. First, teacher preferences differ across tokens, even within the same response. Second, at the token level, majority voting yields a subset of reliable teachers, and the preferred teacher always exhibits a large student-teacher KL distance. The first finding validates our claim, and the second finding motivates a practical rule for token-level teacher selection.

Based on these findings, we propose Multi-Teacher Adaptive On-policy Self-Distillation (MTA-OPSD), the first token-level adaptive OPSD method. As shown in Figure 3, MTA-OPSD constructs a multi-teacher pool consisting of candidate teachers, by varying the truncation strategy and exposure ratio of the privileged reference. During training, rather than using a fixed teacher for all tokens in the same response, each token adaptively selects its own teacher from the multi-teacher pool based on its preference, through majority voting followed by KL-based selection. Our token-level adaptive OPSD design is simple yet effective, precisely addressing the student-teacher mismatch. By assigning each token its preferred teacher, MTA-OPSD provides a more balanced learning signal rather than an overly sharp, rigid teacher distribution from a single teacher. We evaluate MTA-OPSD on mathematical reasoning benchmarks using models from the Qwen3 family (Yang et al., 2025), and it consistently outperforms existing OPSD approaches with an average gain of up to $+5.9$ over the base model.

Our contributions are summarized as follows:

- We quantify the teacher preference across tokens via a teacher-intervention pilot study, validating that preferred teachers can differ from token to token within the same response. We further reveal that the token-level preferred teacher can be identified through majority voting and KL-based selection.
- We propose MTA-OPSD, the first token-level adaptive OPSD method. By constructing a multi-teacher pool as the supervision source, our approach enables each token to adaptively select its preferred teacher under the guidance of majority voting and KL divergence, effectively addressing the student-teacher mismatch.
- We conduct extensive experiments across various models on mathematical reasoning tasks, demonstrating that our approach achieves superior reasoning performance compared to existing OPSD methods. Furthermore, we empirically analyze comprehensive ablation studies, paving the way for future advances in this field.

# 2 Related Work

## 2.1 On-policy Distillation and On-Policy Self-distillation

Knowledge distillation (Hinton et al., 2015) transfers knowledge from a large teacher model to a student model by training on the teacher's soft output distributions. Early approaches (Kim & Rush, 2016; Jiao et al., 2020; Wang et al., 2020) leveraged it to sequence-level distillation, establishing the dominant off-policy distillation approaches. (Gu et al., 2024; Agarwal et al., 2024; Lu & Lab, 2025; Yang et al., 2026b) extended it to On-policy Distillation (OPD), addressing the exposure bias (Bengio et al., 2015) mismatch by shifting the training distribution to the student's own generations.

Despite the rapid progress of OPD, the paradigm relies on a separate, typically stronger teacher model, which may be unavailable or expensive to maintain. To address this challenge, OPSD (Zhao et al., 2026; Hübotter et al., 2026; Shenfeld et al., 2026; Luo et al., 2026; Yang et al., 2026a) removes this requirement by instantiating the teacher and student from the same model. Under this setting, the teacher and student share the same model, but the teacher is granted access to privileged reference unavailable to the student.

However, recent studies (Kim et al., 2026; Wang et al., 2026b; Kaur et al., 2026; Harne et al., 2026; Shen et al., 2026) reveal a persistent mismatch in OPSD. In particular, a highly informative reference can induce an overly sharp teacher distribution and create a mismatch between the teacher's supervision and the student's learning dynamics. These observations motivate the line of research for adaptive OPSD methods.

## 2.2 Adaptive On-Policy Self-distillation

Adaptive OPSD seeks to align the teacher supervision with the student's learning dynamics. Rather than exposing the complete privileged reference, existing methods (Tan & Hong, 2026; Han et al., 2026b; Yang et al., 2026c) control its content or exposure ratio using mechanisms such as learned reference-exposure schedules. By varying the amount of privileged references available to the teacher, these methods aim to avoid overly sharp supervision while retaining the benefits of dense learning signals.

However, these methods adapt at the response level, where all tokens in the response share the same teacher and the same privileged reference. In contrast, we identify that each token needs its own teacher as tokens have distinct teacher preferences (Section 4), and further propose our token-level multi-teacher adaptive OPSD method.

# 3 Preliminaries

OPSD is a special case of OPD. OPD transfers knowledge from a stronger teacher model $p_T$ to a weaker student model $p_\theta$ by enforcing dense supervision over trajectories sampled by the student. Given a prompt $x$, the student samples a response $y = \{y^1, y^2, \cdots, y^L\}$, where $L$ denotes the generation length and $y^i$ refers to the $i$-th token in the response. The learning objective is to minimize the token-level KL between the teacher's and the student's next-token distributions:

$$\begin{aligned}\mathcal{L}_{\text{OPD}}(\theta) &= \mathbb{E}_x \left[ \sum_{i=1}^{L} \mathcal{D}_{\text{KL}} \left( p_\theta \left( \cdot \mid y^{<i}, x \right) \,||\, p_T \left( \cdot \mid y^{<i}, x \right) \right) \right] \\ &= \mathbb{E}_x \left[ \sum_{i=1}^{L} \mathcal{D}_{\text{KL}} \left( p_\theta \left( \cdot \mid c_i \right) \,||\, p_T \left( \cdot \mid c_i \right) \right) \right]\end{aligned} \tag{1}$$

where $p(\cdot \mid y^{<i}, x)$ denotes the distribution over the next token $y^i$, and we let $c_i$ refer to $(y^{<i}, x)$ for simplicity. While forward KL and reverse KL are most used, other distribution divergence measures like generalized Jensen-Shannon divergence (Agarwal et al., 2024) can also be employed.

Recent advances have extended OPD to OPSD, where the student and teacher are instantiated from the same model $p_\theta$. The difference lies entirely in their conditioning contexts. Privileged reference $r$, such as reference solutions (Zhao et al., 2026) or environment feedback (Hübotter et al., 2026), is

appended to the original prompt to construct a teacher-specific prompt. Thus, the teacher distribution is:

$$p_T(\cdot \mid c_i) = p_\theta(\cdot \mid c_i, r). \tag{2}$$

Consequently, the learning objective in Equation (1) adapts into the following:

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_x \left[ \sum_{i=1}^{L} \mathcal{D}_{\text{KL}}\left(p_\theta(\cdot \mid c_i) \,||\, p_\theta(\cdot \mid c_i, r)\right) \right]. \tag{3}$$

In this setting, both the teacher and student share the same model but differ only in the conditioning contexts, and the response is solely generated from the student.

## 4 Pilot Study: Token Matters

Before diving into a detailed design, a critical question must first be answered:

*Given $c_i = (y^{<i}, x)$, how can we determine if one teacher $p_T(c_i)$ is more suitable than another?*

A direct approach would be to train on every token separately with specific teachers, covering enough token-teacher combinations. However, it is obviously not practical.

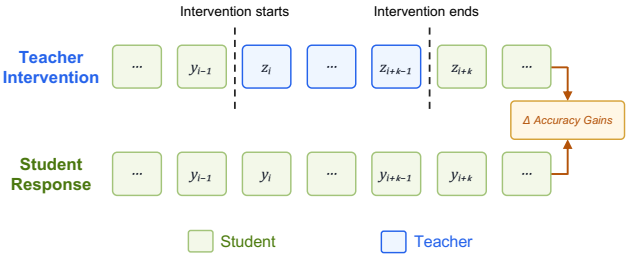


Figure 2: Illustration of the teacher-intervention pilot study.

We instead answer this question from another view. Motivated by (Xu et al., 2025; 2026a), we estimate teacher suitability through teacher intervention. Specifically, given a prompt $x$ and a student's response $y = \{y^1, y^2, \cdots, y^L\}$, we select a token position $1 < i < L$ and preserve the student-generated prefix before that position. Starting from position $i$, a candidate teacher takes over the generation process and produces a short subsequence:

$$\begin{aligned} z_{1:i-1} &= y_{1:i-1} = p_\theta(x) \\ z_{t:t+k} &= p_T(y^{<i}, x) = p_\theta(c_i, r), \end{aligned} \tag{4}$$

where $k$ denotes the teacher-intervention length. After this intervention, control is returned to the student, which completes the remainder of the response:

$$z_{i+k+1:L} = p_\theta(z^{<i+k+1}, x). \tag{5}$$

Thus, we obtain a "student-teacher-student" relay response $z = \{z_1, z_2, \cdots, z_L\}$. We compare its accuracy with the pure student response, which serves as a proxy for the candidate teacher preference at the selected token position. A larger gain indicates a more suitable teacher, whereas a smaller gain indicates the opposite.

The main purpose of this pilot study is that, by varying both the intervention token position and the candidate teacher, this study separates variation along the token and teacher axes. This allows us to examine whether teacher preferences changes across token positions and to identify its patterns. Interestingly, the pattern is simple but clear. We next describe the experimental setup and then present our findings, as highlighted in Section 4.2.

### 4.1 Setup

Our study varies two principal factors: tokens at which the intervention begins, and teachers which are used for the intervention.

For the token axis, we consider both regular-spaced and statistic-extreme token positions. For the former, we select tokens located at 25%, 50%, and 75% of the original student rollout. For the latter,

we select tokens associated with the highest and lowest student entropy, and the largest and smallest student-teacher KL distance [1]. In total, we select 7 intervention positions from each rollout.

For the teacher axis, we vary both the reference-truncation strategy[2] and the reference-exposure ratio. We consider prefix, suffix, and random-span for truncation strategy, each with exposure ratios of $0.5$ and $0.75$, together with a full-reference teacher. This produces 7 candidate teachers.

We evaluate intervention lengths across $k = 64$, $128$, and $256$. We sample $64$ prompts of at least medium difficulty[3] from Deepmath-103k (He et al., 2026). We conduct the study using Qwen3-1.7B and Qwen3-4B as student models. To reduce sampling variance, for every experimental configuration, we sample 2 teacher continuations and then 2 student relay following each continuation. In total, $64 * 7 * 7 * 3 * 2 * 2 * 2 = 75264$ generations are evaluated.

### 4.2 Results and Findings

Interestingly, the results show clear and meaningful findings.

**Findings1:** Teacher preference varies across tokens.
**Findings2:** Majority voting provides a subset of reliable teachers.
**Findings3:** The preferred teacher always exhibit a large KL distance from the student.

**Teacher preference varies by tokens.** We define the preferred teacher at a token position as the candidate that produces the largest intervention gain. To determine whether this preference changes within a response, we measure how often the preferred teacher is shared by different regular-spaced tokens in the same rollout. As shown in Table 1, for Qwen3-1.7B with $k = 256$, the overlap is only approximately 50%, which indicates that the teacher preference differs by tokens. These reulsts directly support our claim: every token needs its own teacher.

Table 1: Overlap of preferred teachers between tokens within the same rollout.

| Token Position | 0.25 | 0.50 | 0.75 |
|---|---|---|---|
| **0.25** | - | 48.44% | 50.00% |
| **0.50** | 48.44% | - | 51.56% |
| **0.75** | 50.00% | 51.56% | - |

**Majority voting identifies reliable teachers.** By comparing the teacher distribution with the student one over a student-generated token $y^i$, we can obtain the advantage signal:

$$a_T(y^i \mid c_i) = \log\left(p_T\left(y^i \mid c_i\right)\right) - \log\left(p_\theta\left(y^i \mid c_i\right)\right) = \log\frac{p_T(y^i \mid c_i)}{p_\theta(y^i \mid c_i)}. \tag{6}$$

$a_T(\cdot \mid c_i)$ indicates the teacher's relative support for the student's choice $y^i$. Specifically, If $a_T(\cdot \mid c_i) > 0$ means that the teacher favors the sampled token; otherwise, it indicates that the student's choice is relatively opposed by the teacher. Thus, at each token position, we can divide the candidate teachers into two groups based on this sign, and obtain the majority subset.

Table 2: Fraction of preferred teachers contained in the majority subset.

| Model \ Token Position | Regular | Extreme |
|---|---|---|
| **Qwen3-1.7B** | 77.94% | 88.09% |
| **Qwen3-4B** | 87.11% | 90.68% |

Table 3: Average gains of teachers in the majority and minority subsets.

| Group | Majority | Minority | Overall |
|---|---|---|---|
| **Gain** | +3.34% | +1.88% | +2.34% |

We then examine whether the teacher producing the largest gain belongs to the majority subset. As shown in Table 2, in around $80\% - 90\%$ cases it falls within this subset. We

[1] For identifying these KL-based positions, the teacher is conditioned on the full reference.

[2] The truncation strategy determines which part of the reference is retained. For example, prefix truncation with an exposure ratio of $0.5$ retains the first half of the reference. The final answer is retained under every truncation setting to ensure a fair comparison.

[3] For each prompt, we first sample $4$ student responses and exclude prompts for which all four responses are correct.

further compare the average gain of teachers in both the majority and minority subsets. For Qwen3-4B with $k = 256$, as shown in Table 3, the majority subset always achieves larger gains per teacher. These results indicate that majority voting provides a reliable first-stage filter for teacher selection.

**Large KL distinguishes the preferred teacher.** We also found interesting correlation between the teacher preference and the student-teacher KL distance. For each token, we select the teachers with the largest and smallest student-teacher KL distance from the majority subset and compare their gains. As shown in Table 4, in regular token positions, the farthest teacher consistently produces greater average gains. This observation suggests that, student-teacher KL distance provides a useful criterion for teacher preference, for each token.

Table 4: Average gains from the farthest and nearest teachers.

| Model \ Teacher | farthest | Nearest | $\Delta$ |
|---|---|---|---|
| **Qwen3-1.7B** | +2.42% | +0.59% | +1.83% |
| **Qwen3-4B** | +2.78% | +1.21% | +1.57% |

# 5 METHOD

Based on the findings in Section 4.2, we propose Multi-Teacher Adaptive OPSD (MTA-OPSD), a simple yet effective framework for token-level adaptive OPSD. As illustrated in Figure 3, rather than using a single teacher, MTA-OPSD constructs a multi-teacher pool consisting of candidate teachers conditioned on different variants of the privileged reference. Through majority voting followed by KL-based selection, it assigns each token its preferred teacher. This ensures that different tokens within the same response receive supervision from their respective teachers.

## 5.1 MULTI-TEACHER POOL

To support token-level adaptive teacher selection, we construct a multi-teacher pool $M$, consisting of all candidate teachers available for each token:

$$M = \{p_{T_1}, p_{T_2}, \cdots, p_{T_K}\}, \tag{7}$$

Where $K$ is the number of candidate teachers and $p_{T_k}$ denotes the $k$-th teacher. All teachers share the same model $p_\theta$ with the student. They differ from each other only in the privileged references on which they are conditioned.:

$$R = \{r_1, r_2, \cdots, r_K\}, \tag{8}$$

which leads to the next-token distribution of each teacher as:

$$p_{T_k} = p_\theta(\cdot \mid x, r_k). \tag{9}$$

Each $r_k$ differs from each other by its truncation way and exposure ratio of the full reference. An example of our multi-teacher pool is provided in Appendix F.

While the multi-teacher pool supplies a diverse set of supervision signals, it does not determine which teacher should supervise each token. Guided by the pilot study in Section 4, MTA-OPSD performs token-level adaptive teacher assignment in two steps: (1) majority voting identifies a reliable subset of teachers, and (2) KL-based selection chooses the preferred teacher from that subset.

## 5.2 MAJORITY VOTING

As shown in Section 4.2, majority voting provides a reliable teacher subset by identifying the direction of the advantage signals. We therefore use majority voting as the first step of teacher selection.

Specifically, for a token $y^i$ in the student response, the multi-teacher pool produces the following set of its teacher token distributions:

$$\begin{aligned} P_M(c_i) &= \{p_{T_1}(\cdot \mid c_i), p_{T_2}(\cdot \mid c_i), \cdots, p_{T_K}(\cdot \mid c_i)\} \\ &= \{p_\theta(\cdot \mid c_i, r_1), p_\theta(\cdot \mid c_i, r_2), \cdots, p_\theta(\cdot \mid c_i, r_K)\}. \end{aligned} \tag{10}$$

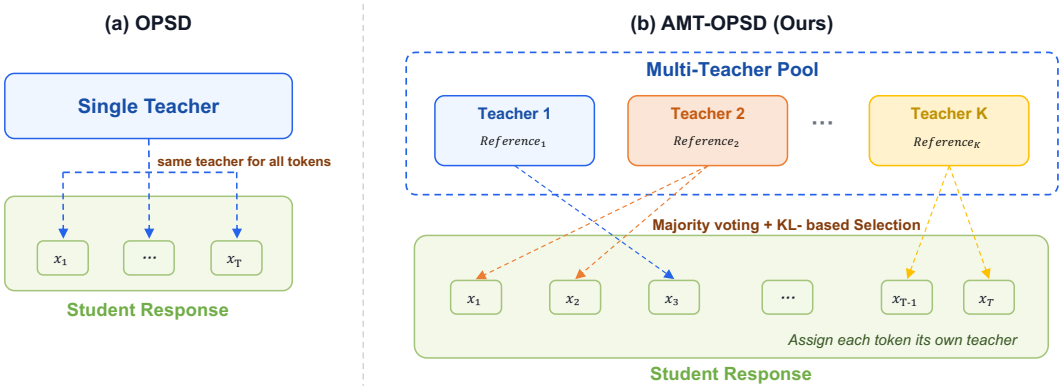


Figure 3: The framework of our approach, MTA-OPSD. With the multi-teacher pool, it adaptively assigns each token its preferred teacher, by majority voting followed by KL-based selection.

For each teacher token distribution in $p_M$, we compare it with the student token distribution $p_\theta(\cdot \mid y^{<i}, x)$ over $y^i$, to obtain the following set:

$$A_M(c_i) = \{a_{T_1}(y^i \mid c_i), a_{T_2}(y^i \mid c_i), \cdots, a_{T_K}(y^i \mid c_i)\}, \tag{11}$$

where advantage $a_{T_K}(y^i \mid c_i)$ is computed using Equation (6). The sign of $a_{T_k}(y^i \mid c_i)$ indicates the $k$-th teacher's relatively support for the student's choice at $y^i$. We denote it as the supporting binary vote:

$$v_{T_k}(c_i) = \operatorname{sgn}\left(a_{T_k}\left(y^i \mid c_i\right)\right) = \begin{cases} +1, & a_{T_k}\left(y^i \mid c_i\right) > 0, \\ -1, & a_{T_k}\left(y^i \mid c_i\right) < 0. \end{cases} \tag{12}$$

We then determine the majority voting direction at token position $i$:

$$V_M(c_i) = \operatorname{sgn}\left(\sum_{k=1}^{K} v_{T_k}\left(c_i\right)\right), \tag{13}$$

and the corresponding token-level majority teacher subset[4]:

$$M_{\text{maj}}(c_i) = \left\{p_{T_k} \in M \mid v_{T_k}\left(c_i\right) = V_M\left(c_i\right)\right\}. \tag{14}$$

Note that this procedure is token-level, performing independently at every token position.

### 5.3 KL-BASED SELECTION

Majority voting filters out teachers whose supervision direction disagrees with most candidates, but it does not distinguish among teachers within the majority subset. As shown in Section 4.2, the preferred teacher always has large KL distance from the student. We therefore use KL distance as the second-stage selection criterion.

Specifically, for token $y^i$ in the student response and its majority teacher subset $M_{\text{maj}}(c_i)$, the token-level preferred teacher is defined as:

$$p_T^*(c_i) = p_\theta(\cdot \mid c_i, r^*) = \arg \max_{p_{T_k} \in M_{\text{maj}}(c_i)} \mathcal{D}_{\text{KL}}\left(p_\theta\left(\cdot \mid c_i\right) \,||\, p_{T_k}\left(\cdot \mid c_i\right)\right). \tag{15}$$

Note that since the preferred teacher is determined separately for every token, it can vary across tokens within the same response. Consequently, Substituting this token-level teacher into the learning objective in Equation (3) gives:

$$\begin{aligned} \mathcal{L}_{\text{MTA-OPSD}}(\theta) &= \mathbb{E}_x\left[\sum_{i=1}^{L} \mathcal{D}_{\text{KL}}\left(p_\theta\left(\cdot \mid c_i\right) \,||\, p_T^*\left(\cdot \mid c_i\right)\right)\right] \\ &= \mathbb{E}_x\left[\sum_{i=1}^{L} \mathcal{D}_{\text{KL}}\left(p_\theta\left(\cdot \mid c_i\right) \,||\, p_\theta\left(\cdot \mid c_i, r^*\right)\right)\right], \end{aligned} \tag{16}$$

Thus, MTA-OPSD assigns each token its own teacher.

[4]If a tie occurs between the majority and minority votes, we then compare which side has the larger absolute value. Additional details are provided in Appendix B

Table 5: Mathematical reasoning performance across model scales. MTA-OPSD achieves the best average performance at all three model scales.

| Model | Method | AMC23 | AIME24 | AIME25 | Average |
|---|---|---|---|---|---|
| Qwen3-1.7B | Base Model | 74.1 | 45.6 | 33.6 | 51.1 |
| | OPSD | 75.3 | 49.2 | 37.5 | 54.0 |
| | SDPO | 73.5 | 43.3 | 31.4 | 49.4 |
| | Paint | 76.0 | 47.5 | 38.6 | 54.0 |
| | **MTA-OPSD** | 78.5 | 51.4 | 41.1 | $57.0_{+5.9}$ |
| Qwen3-4B | Base Model | 88.6 | 73.1 | 65.3 | 75.7 |
| | OPSD | 89.1 | 74.4 | 65.6 | 76.4 |
| | SDPO | 87.7 | 71.4 | 61.7 | 73.6 |
| | Paint | 89.1 | 71.7 | 66.1 | 75.6 |
| | **MTA-OPSD** | 89.0 | 74.4 | 68.1 | $77.2_{+1.5}$ |
| Qwen3-8B | Base Model | 89.1 | 73.1 | 66.9 | 76.4 |
| | OPSD | 89.9 | 75.0 | 67.8 | 77.6 |
| | SDPO | 88.9 | 73.9 | 65.6 | 76.1 |
| | Paint | 90.0 | 76.1 | 68.3 | 78.9 |
| | **MTA-OPSD** | 89.8 | 77.5 | 70.3 | $79.2_{+2.8}$ |

# 6 Experiments

## 6.1 Experiment Setup

Additional experiment setup details are provided in Appendix C.

**Models and Datasets.** We evaluate models from the Qwen3 family (Yang et al., 2025) at three scales: Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, all with thinking mode enabled during both training and validation. We conduct experiments across mathematical reasoning tasks. The training data is drawn from DeepMath-103K (He et al., 2026), and we evaluate on AMC23(MAA, 2023), AIME24(Zhang & Math-AI, 2024), and AIME25(Zhang & Math-AI, 2025).

**Baselines.** We compare our method against two fixed-teacher OPSD methods and one adaptive OPSD method: (1) OPSD (Zhao et al., 2026), which performs dense-logit distillation using forward KL with token-wise pointwise KL clipping; (2) SDPO (Hübotter et al., 2026), which performs dense-logit distillation using Jensen–Shannon divergence; and (3) Paint (Tan & Hong, 2026), which adaptively determines the reference-exposure ratio according to the student-teacher gap.

**Implementation Details.** All baselines and our method use dataset-provided ground-truth reference solutions as the privileged information, and all contexts include both the reasoning trace and the final answer. We use identical training and evaluation settings across all methods, with maximum generation lengths of 4096 tokens during training and 38912 tokens during validation. The multi-teacher pool contains seven teacher configurations:{full reference, prefix-0.5,prefix-0.75,suffix-0.5,suffix-0.75,random-span-0.5,random-span-0.75}[5]. We train each method for 300 steps, evaluate it every 50 steps, and report its best Avg@12 results.

## 6.2 Main Results

Table 5 presents a comprehensive reasoning performance comparison between existing OPSD methods and our approach. For average performance, MTA-OPSD achieves the highest average performance at every model scale, yielding average improvement of $+5.9$ for Qwen3-1.7B, $+1.5$ for Qwen3-4B, and $+2.8$ for Qwen3-8B, over the base model. These results outperform existing OPSD methods. For individual benchmark, MTA-OPSD also achieves or matches the best result on nearly every model-benchmark combination. Note that the improvements are particularly consistent and strong on the more challenging AIME25 benchmark. These results highlight the importance of token-level adaptive teacher assignments. An qualitative comparison is provided in Appendix D.

[5]The truncation strategy determines which part of the reference is retained. For example, prefix truncation with an exposure ratio of $0.5$ retains the first half of the reference. The final answer is retained under every truncation setting to ensure a fair comparison.

### 6.3 Ablation Studies

All ablation studies use Qwen3-1.7B as the base model and follow the same experiment settings.

**Effect of the Multi-Teacher Pool.** We first investigate whether the multi-teacher pool provides more effective supervision than any individual teacher. In Table 6, we compare our method with training using each teacher configuration from the pool independently. Although the effectiveness of individual teachers varies from each other, MTA-OPSD achieves the highest score on both average and individual benchmarks. These results suggest that the multi-teacher pool offers diverse, sufficient, and complementary learning signals.

Table 6: Comparison of MTA-OPSD with single-teacher variants.

| Teacher | AMC23 | AIME24 | AIME25 | Average |
|---|---|---|---|---|
| Base Model | 74.1 | 45.6 | 33.6 | 51.1 |
| Full (OPSD) | 75.3 | 49.2 | 37.5 | 54.0 |
| Prefix-0.5 | 76.8 | 50.8 | 40.0 | 55.9 |
| Prefix-0.75 | 78.2 | 49.4 | 41.1 | 56.2 |
| Suffix-0.5 | 75.9 | 48.6 | 34.7 | 53.1 |
| Suffix-0.75 | 75.0 | 50.3 | 40.0 | 55.1 |
| Random-span-0.5 | 76.8 | 49.4 | 36.9 | 54.4 |
| Random-span-0.75 | 76.4 | 50.6 | 35.8 | 54.3 |
| **MTA-OPSD** | 78.5 | 51.4 | 41.1 | 57.0 |

Table 7: Comparison of token-level teacher-assignment strategies using the same multi-teacher pool.

| Strategy | AMC23 | AIME24 | AIME25 | Average |
|---|---|---|---|---|
| Base Model | 74.1 | 45.6 | 33.6 | 51.1 |
| Random-pick | 76.1 | 49.2 | 35.5 | 53.6 |
| Mean-loss | 75.7 | 48.9 | 36.4 | 53.7 |
| **MTA-OPSD** | 78.5 | 51.4 | 41.1 | 57.0 |

**Effect of Token-Level Teacher Assignment.** The multi-teacher pool is accompanied with the token-level adaptive teacher assignment. The previous ablation establishes the value of the former, but does not isolate the contribution of the latter. We therefore retain the same multi-teacher pool and compare MTA-OPSD with two alternative assignment strategies: (1) Random teacher selection from the pool for each token, and (2) averaging the distillation losses from all teachers. As shown in Table 7, MTA-OPSD consistently outperforms both alternatives on every benchmark. These show that the gains do not arise solely from multiple teachers; selecting a suitable teacher for each token is critical to using the teacher pool effectively.

Table 8: Ablation of majority voting and KL-based teacher selection.

| Majority Voting | KL Selection | AMC23 | AIME24 | AIME25 | Average |
|---|---|---|---|---|---|
| No | Farthest | 75.8 | 48.9 | 33.9 | 52.9 |
| **Yes** | **Farthest** | 78.5 | 51.4 | 41.1 | 57.0 |
| | Random | 78.0 | 51.1 | 40.8 | 56.6 |
| | Nearest | 75.7 | 48.6 | 35.6 | 53.3 |

**Teacher-Assignment Dynamics.** We also investigate how frequently each teacher is selected throughout training. The assignments are distributed relatively evenly across the teacher pool, indicating that the selection mechanism does not collapse to a single dominant teacher. We also observe that, under the same truncation strategy, a higher reference-exposure ratio does not necessarily produce a larger student-teacher KL distance, as discussed in Appendix E. This observation further suggests that existing adaptive methods of adjusting reference exposure alone cannot determine teacher suitability.

**Effect of Majority Voting and KL-Based Selection.** MTA-OPSD selects a teacher in two steps: majority voting and KL-based selection. Table 8 presents the contribution of each step, where the second row is our default setting. MTA-OPSD consistently outperformes all variants, further validating the effectiveness of our design: majority voting first identifies a reliable teacher subset, and KL-based selection chooses a teacher from that subset.

## 7 Conclusion

We present MTA-OPSD, the first token-level adaptive OPSD method. A teacher-intervention study shows that teacher preferences vary across tokens and motivates our two-step token-level adaptive teacher selection strategy: majority voting with KL-based selection. Future work may explore self-evolving teacher pools and more stable multi-teacher training.

### AI USE STATEMENT

In this work, we used generative AI tools for [design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, interpret results]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses], and [Generate synthetic data sets, clean and reformat dataset] are not applicable to this work. Additionally, we used generative AI tools for [edit a research paper to improve readability]. We have reviewed all AI-assisted work. [we checked LLM-polished manuscript and verified LLM-generated code]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We have uploaded our code in the supplementary material, providing all configures and parameters needed for reproducing our work. The experiment setup and hardware requirement are provided in Section 6.1 and Appendix C.

## 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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Fangming Cui, Sunan Li, and Jiahong Li. A brief overview: On-policy self-distillation in large language models. *arXiv preprint arXiv:2605.18141*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

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.

Zhuowen Han, Jinwei Xiao, Zhengxi Lu, Renren Jin, Zhiyuan Yao, Yuxin Liu, Hongyan Hao, Yueqing Sun, Yu Yang, Qi Gu, et al. Distill where you fail: Recovering learning signals of negative rl-groups from adaptive teacher guidance. *arXiv preprint arXiv:2608.00782*, 2026a.

Zihao Han, Tiangang Zhang, Huaibin Wang, and Yilun Sun. Adaptive teacher exposure for self-distillation in llm reasoning. *arXiv preprint arXiv:2605.11458*, 2026b.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in rlvr: An entropy change perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31105–31133, 2026.

Sarthak Harne, Chinmay Karkar, Yash Pandya, Ahmed Awadallah, and Akshay Nambi. Privileged, but biased: How pi-conditioned teachers break self-distillation. *arXiv preprint arXiv:2608.04794*, 2026.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 138306–138322, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 4163–4174, 2020.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. *arXiv preprint arXiv:2605.13255*, 2026.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *arXiv preprint arXiv:2603.24472*, 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.

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

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

Yifu Luo, Zeyu Chen, Haoyu Wang, Xinhao Hu, Yuxuan Zhang, Zhizhou Sha, and Shiwei Liu. Learning from the self-future: On-policy self-distillation for dllms. *arXiv preprint arXiv:2606.18195*, 2026.

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

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.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, et al. Purified opsd: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Zhiquan Tan and Yinrong Hong. Paint: Partial-solution adaptive interpolated training for self-distilled reasoners. *arXiv preprint arXiv:2604.26573*, 2026.

Jiaxuan Wang, Xuan Ouyang, Zhiyu Chen, Yulan Hu, Zheng Pan, Xin Li, and Lan-Zhe Guo. Trace: Distilling where it matters via token-routed self on-policy alignment. *arXiv preprint arXiv:2605.10194*, 2026a.

Meng Wang, Haohan Zhao, Wenzhuo Liu, Lu Yang, Geng Liu, Haiyang Guo, Guo-Sen Xie, Gaofeng Meng, Hongbin Liu, and Fei Zhu. Denser $\neq$ better: Limits of on-policy self-distillation for continual post-training. *arXiv preprint arXiv:2607.01763*, 2026b.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:115452–115486, 2026c.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Zhenlin Wei, Pu Jian, Yingzhuo Deng, Xiaohan Wang, Jiajun Chai, Zhexin Hu, Wei Lin, Shanbin Zhang, and Guojun Yin. Amr-sd: Asymmetric meta-reflective self-distillation for token-level credit assignment. *arXiv preprint arXiv:2605.18529*, 2026.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation. *arXiv preprint arXiv:2607.26057*, 2026a.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

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*, 2026b.

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.

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.

Yongkang Yang, Zhezheng Hao, Hong Zhang, Yi Liu, Xiankun Lin, Wence Ji, Fanjunduo Wei, Jiarui Yu, Qiang Lin, Xiaoyun Liang, et al. Matching supervision to the student's learning capacity: A unified framework for on-policy self-distillation. *arXiv preprint arXiv:2608.08176*, 2026c.

Ruixiang Zhang, Richard He Bai, Huangjie Zheng, Navdeep Jaitly, Ronan Collobert, and Yizhe Zhang. Embarrassingly simple self-distillation improves code generation. *arXiv preprint arXiv:2604.01193*, 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 VISUALIZATION DETAILS

The three teachers in Figure 1 are with configurations of prefix-0.5, suffix-0.5, and random-span-0.5. The entire student response is shown in Figure 4.

*Student Response*

```
<think>
Okay, so I need to evaluate this line integral: ∫γ 1/(z - a) dz, where γ is the contour γ = a
+ Re^{it} for t from 0 to 2π. And a is a complex number. Hmm, let me think about how to
approach this.

First, I remember that line integrals over complex functions can sometimes be evaluated using
the residue theorem. But before jumping into that, maybe I should try parametrizing the
integral and see if I can compute it directly.

So, the contour γ is given by z = a + Re^{it}, where t goes from 0 to 2π. Wait, R is a real
number, right? So this is a circle centered at a with radius R. So, the contour γ is a circle
of radius R centered at the point a in the complex plane. But a is a complex number, so the
center of the circle is a, and the radius is R.

•••

Substitute into the integral:

$$
\int_{\gamma} \frac{1}{z - a} \, dz = \int_0^{2\pi} \frac{1}{Re^{it}} \cdot iRe^{it} \, dt =
\int_0^{2\pi} i \, dt = 2\pi i.
$$

This confirms the result obtained via the residue theorem.

---

### Final Conclusion

The integral evaluates to:

$$
\boxed{2\pi i}
$$
```

Figure 4: Entire student response for Figure 1.

## B ADDITIONAL MAJORITY VOTING DETAILS

As discussed in Section 5.2, a tie may occur when the multi-teacher pool contains an even number of teachers, with equal numbers of positive and negative votes. In this case, we break the tie according to the magnitudes of the corresponding advantage signals. Specifically, the majority direction is defined as

$$V_M(c_i) = \begin{cases} \operatorname{sgn}\left(\sum_{k=1}^{K} v_{T_k}(c_i)\right), & \text{if } \sum_{k=1}^{K} v_{T_k}(c_i) \neq 0, \\ \operatorname{sgn}\left(\sum_{k=1}^{K} v_{T_k}(c_i)\left|a_{T_k}(y^i \mid c_i)\right|\right), & \text{otherwise.} \end{cases} \tag{17}$$

Thus, when the numbers of positive and negative votes are equal, we select the group with the larger total absolute advantage. The majority teacher subset is then still constructed using Equation (14). This tie-breaking rule retains the direction supported by the stronger aggregate teacher signal.

## C Additional Experiment Settings

All experiments are conducted on 4 H200 GPUs. We use a learning rate of $1 \times 10^{-6}$. All methods are implemented on top of the veRL (Sheng et al., 2025) PPO trainer with FSDP2 actor sharding and vLLM rollouts. All runs use full-parameter training (no LoRA (Hu et al., 2021)).

Note that following (Pan et al., 2026), we use training problems at difficulty levels 5–7 for Qwen3-1.7B, 6–8 for Qwen3-4B, and levels 7–10 for Qwen3-8B.

## D Quality Visualization

We provide a qualitative comparison in Figure 5.

*Question: Sixteen chairs are arranged in a row. Eight people each select a chair in which to sit so that no person sits next to two other people. Let $ N $ be the number of subsets of 16 chairs that could be selected. Find the remainder when $ N $ is divided by 1000. Please reason step by step, and put your final answer within \boxed{}. Assistant*

*Ground Truth: 907 (N = 2907)*

*Model: AMT-OPSD (Ours)*

```
To solve this problem, we need to determine the number of subsets of 16 chairs that can be
selected such that **no person is seated between two others**. This is equivalent to
ensuring that **no three consecutive chairs are occupied**, since if three people are
seated consecutively, the middle one is flanked by two others, violating the condition.

--- ### Reformulating the Problem

We are to choose **8 chairs out of 16** such that **no three chairs are occupied
consecutively**. We are to choose **8 chairs out of 16** such that **no three chairs are
occupied consecutively**. This is a classic combinatorial problem involving **binary
strings of length 16 with exactly 8 ones**, where **no three consecutive ones are
allowed**.

•••

--- ### Final Step:
Compute Modulo 1000 We are asked to find the remainder when $ N = 2907 $ is divided by
1000.

$$ 2907 \mod 1000 = \boxed{907} $$

--- ### Final Answer

$$ \boxed{907} $$
```

*Model: OPSD*

```
To solve the problem, we are given a row of 16 chairs and asked to determine how many
valid subsets of 8 chairs (i.e., seating arrangements of 8 people) can be selected such
that **no person sits next to two other people**.

--- ### Step 1: Understanding the Constraint

The key constraint is that **no person can sit next to two other people**. This implies
that **each person must be seated in a way that at least one empty chair separates them
from any other seated person**.

•••

--- ### Step 3: Final Computation

Now, the problem asks for the remainder when the number of such subsets is divided by
1000. Since the number of subsets is **9**, the remainder when divided by 1000 is simply:

 $$ 9 \mod 1000 = 9 $$

--- ### Final Answer

$$ \boxed{9} $$
```

Figure 5: Quality visualization.

## E Additional Details about KL-distance

As discussed in Section 6.3, even under same truncation strategy, larger reference-exposure ratio does not lead to larger student-teacher KL distance. In our pilot study Section 4.2, for example, for Qwen3-4B modle with suffix truncation, only $49.1\%$ cases that $0.7$ exposure ratio has larger KL distance from the student. In contrast, the data for $0.5$ exposure ratio is $50.9\%$.

## F Multi-teacher Pool Visualization

Figures below demonstrate a brief visualization of our multi-teacher pool.

*Teacher o (Full )*

```
<|im_start|>user
Problem: Find the auxiliary equation for the ordinary differential equation with constant
coefficients: \((x^2D^2 + xD + 1)y = \sin(2\log x)\sin(\log x)\).

Here is a reference solution to this problem:
=== Reference Solution Begin ===
Okay, so I need to find the auxiliary equation for the given ordinary differential equation
(ODE) with constant coefficients: \((x^2D^2 + xD + 1)y = \sin(2\log x)\sin(\log x)\). Hmm,
let's see. First, I remember that when dealing with linear ODEs with variable coefficients,
especially ones that have terms like \(x^2D^2\) and \(xD\), a common technique is to use a
substitution to transform the equation into one with constant coefficients. The substitution
usually involves changing the independent variable from \(x\) to another variable, say \(t\),
which is a logarithm function of \(x\).

...

y involves changing the independent variable from \(x\) to another variable, say \(t\), which
is a logarithm function of \(x\).

Let me recall the exact substitution. Since we have terms like \(x^n D^n\), this makes me
think of the Cauchy-Euler equation, which typically has the form \(x^n y^{(n)} + \dots + a_1
x y' + a_0 y = f(x)\). For such equations, the substitution \(t = \log x\) (or sometimes \(t
= \ln x\)) is used to convert the equation into a linear ODE with constant coefficients. That
seems to fit here because the given equation has \(x^2D^2\), \(xD\), and constant coefficient
1. So, let's try that substitution.

Let me set \(t = \log x\). Then, \(x = e^t\). Now, I need to express the differential
operators \(D = \frac{d}{dx}\) in terms of \(t\).

...

But the question asks for the auxiliary equation of the given ODE. Wait, the original ODE is
nonhomogeneous, but the auxiliary equation is associated with the homogeneous equation. The
auxiliary equation itself is just the characteristic equation obtained from the homogeneous
part. So even though the right-hand side has sines and cosines, the auxiliary equation is
determined solely by the left-hand side after substitution.

So, after substitution, the left-hand side becomes \(y'' + y\), so the auxiliary equation is
\(m^2 + 1 = 0\). Therefore, the answer should be \(m^2 + 1 = 0\).

But wait, let me make sure I didn't skip any steps. The original equation is in terms of
\(x\), but after substitution, it's transformed into a constant coefficient equation in terms
of \(t\).

...

Therefore, the homogeneous part is \(y'' + y = 0\), so the auxiliary equation is \(m^2 + 1 =
0\). Thus, the answer should be \(m^2 + 1 = 0\), which is the auxiliary equation.

But let me check if the original problem was written correctly. The user specified "the
ordinary differential equation with constant coefficients". However, the given equation has
variable coefficients because of the \(x^2 D^2\) and \(xD\) terms.

...

3. **Resulting ODE**: The transformed ODE is \(\frac{d^2 y}{dt^2} + y = \frac{1}{2}[\cos(t) -
\cos(3t)]\).

4. **Auxiliary Equation**: The homogeneous part of the transformed ODE is \(\frac{d^2
y}{dt^2} + y = 0\), leading to the auxiliary equation \(m^2 + 1 = 0\).

Thus, the auxiliary equation for the given differential equation is \(\boxed{m^2 + 1 = 0}\).

Correct final answer: m^2 + 1 = 0
=== Reference Solution End ===

After reading the reference solution above, make sure you understand the reasoning behind
each step.

Please reason step by step, and put your final answer within \boxed{}.<|im_end|>
<|im_start|>assistant
```

Figure 6: Multi-teacher pool visualization.

*Teacher 1 (prefix-0.5 )*

```
<|im_start|>user
Problem: Find the auxiliary equation for the ordinary differential equation with constant
coefficients: \((x^2D^2 + xD + 1)y = \sin(2\log x)\sin(\log x)\).

Here is a reference solution to this problem:
=== Reference Solution Begin ===
Okay, so I need to find the auxiliary equation for the given ordinary differential equation
(ODE) with constant coefficients: \((x^2D^2 + xD + 1)y = \sin(2\log x)\sin(\log x)\). Hmm,
let's see. First, I remember that when dealing with linear ODEs with variable coefficients,
especially ones that have terms like \(x^2D^2\) and \(xD\), a common technique is to use a
substitution to transform the equation into one with constant coefficients. The substitution
usually involves changing the independent variable from \(x\) to another variable, say \(t\),
which is a logarithm function of \(x\).

...

Now, the auxiliary equation for the homogeneous equation \(y'' + y = 0\) is found by
replacing \(y''\) with \(m^2\) and \(y\) with 1, giving \(m^2 + 1 = 0\). Solving this, the
roots are \(m = \pm i\). So, the homogeneous solution is \(y_h = C_1 \cos t + C_2 \sin t\).

But the question asks for the auxiliary equation of the given ODE. Wait, the original ODE is
nonhomogeneous, but the auxiliary equation is associated with the homogeneous equation. The
auxiliary equation itself is just the characteristic equation obtained from the homogeneous
part. So even though the right-hand side has sines and cosines, the auxiliary equation is
determined solely by the left-hand side after substitution.

So, af
... [reference reasoning truncated on the right] ...

Correct final answer: m^2 + 1 = 0
=== Reference Solution End ===

After reading the reference solution above, make sure you understand the reasoning behind
each step.

Please reason step by step, and put your final answer within \boxed{}.<|im_end|>
<|im_start|>assistant
```

Figure 7: Multi-teacher pool visualization.

*Teacher 2 (prefix-0.75 )*

```
<|im_start|>user
Problem: Find the auxiliary equation for the ordinary differential equation with constant
coefficients: \((x^2D^2 + xD + 1)y = \sin(2\log x)\sin(\log x)\).

Here is a reference solution to this problem:
=== Reference Solution Begin ===
Okay, so I need to find the auxiliary equation for the given ordinary differential equation
(ODE) with constant coefficients: \((x^2D^2 + xD + 1)y = \sin(2\log x)\sin(\log x)\). Hmm,
let's see. First, I remember that when dealing with linear ODEs with variable coefficients,
especially ones that have terms like \(x^2D^2\) and \(xD\), a common technique is to use a
substitution to transform the equation into one with constant coefficients. The substitution
usually involves changing the independent variable from \(x\) to another variable, say \(t\),
which is a logarithm function of \(x\).

...

\(x^2 D^2 y + x D y + y = (y'' - y') + y' + y = y'' + y\). So yes, that's correct. Therefore,
the transformed equation is \(y'' + y = \sin(2t)\sin(t)\), which simplifies to \(y'' + y =
[\cos t - \cos 3t]/2\).

Therefore, the homogeneous part is \(y'' + y = 0\), so the auxiliary equation is \(m^2 + 1 =
0\). Thus, the answer should be \(m^2 + 1 = 0\), which is the auxiliary equation.

But let me check if the original problem was written correctly. The user specified "the
ordinary differential equation with constant coefficients". However, the given equation has
variable coefficients because of the \(x^2 D^2\) and \(xD\) terms. B
... [reference reasoning truncated on the right] ...

Correct final answer: m^2 + 1 = 0
=== Reference Solution End ===

After reading the reference solution above, make sure you understand the reasoning behind
each step.

Please reason step by step, and put your final answer within \boxed{}.<|im_end|>
<|im_start|>assistant
```

Figure 8: Multi-teacher pool visualization.

*Teacher 3 (suffix-0.5 )*

```
<|im_start|>user
Problem: Find the auxiliary equation for the ordinary differential equation with constant
coefficients: \((x^2D^2 + xD + 1)y = \sin(2\log x)\sin(\log x)\).

Here is a reference solution to this problem:
=== Reference Solution Begin ===
... [reference reasoning truncated on the left] ...
ter substitution, the left-hand side becomes \(y'' + y\), so the auxiliary equation is \(m^2
+ 1 = 0\). Therefore, the answer should be \(m^2 + 1 = 0\).

But wait, let me make sure I didn't skip any steps. The original equation is in terms of
\(x\), but after substitution, it's transformed into a constant coefficient equation in terms
of \(t\). The auxiliary equation is derived from the homogeneous part of the transformed
equation, which is \(y'' + y = 0\), leading to \(m^2 + 1 = 0\). Therefore, the auxiliary
equation is indeed \(m^2 + 1 = 0\).

But let me double-check the substitution steps to ensure I didn't make a mistake. Let's
rederive \(x^2 D^2 y\) in terms of \(t\).

...

3. **Resulting ODE**: The transformed ODE is \(\frac{d^2 y}{dt^2} + y = \frac{1}{2}[\cos(t) -
\cos(3t)]\).

4. **Auxiliary Equation**: The homogeneous part of the transformed ODE is \(\frac{d^2
y}{dt^2} + y = 0\), leading to the auxiliary equation \(m^2 + 1 = 0\).

Thus, the auxiliary equation for the given differential equation is \(\boxed{m^2 + 1 = 0}\).

Correct final answer: m^2 + 1 = 0
=== Reference Solution End ===

After reading the reference solution above, make sure you understand the reasoning behind
each step.

Please reason step by step, and put your final answer within \boxed{}.<|im_end|>
<|im_start|>assistant
```

Figure 9: Multi-teacher pool visualization.