Demystifying On-Policy Distillation: Roles, Pathologies, and Regulations

Rui Wang[†], **Hongru Wang**[†], **Yi Chen**, **Boyang Xue**[†]
Tianqing Fang[‡*], **Wenhao Yu**[‡], **Kam-Fai Wong**[†*]
[†]The Chinese University of Hong Kong [‡]Tencent AI Lab
ruiwangnlp@outlook.com fangtq229@gmail.com kfwong@cuhk.edu.hk

Abstract

On-policy distillation (OPD) has become a key paradigm in LLM post-training, yet its training dynamics remain poorly understood. We present a systematic study examining the **role**, **pathologies**, and **regulations** of OPD. We first clarify the *role* of OPD as an *exploration catalyst*: it steers the student toward correct reasoning paths via dense token-level guidance, without expanding capability ceiling. We confirm this by showing that prompt diversity matters more than per-problem sampling numbers, and critically, that the effectiveness of OPD hinges entirely on the quality of its guiding signal. This dependency exposes two *pathologies* that derail exploration. The Student-Teacher Mismatch occurs when a large teacher-student distributional gap causes the guiding signal to misalign with task correctness, steering exploration in counterproductive directions. Length Exploitation arises when the aggregated token-level objective creates length-dependent shortcuts, allowing the student to game the reward landscape through response truncation or redundant padding, exploring degenerate length modes rather than reasoning strategies. To tame these pathologies, we investigate lightweight signal *regulations*: *advantage clipping* and *log-scale compression*, ensuring exploration is guided by faithful signals. Experiments across seven benchmarks demonstrate that these regulations eliminate length exploitation and enable effective distillation, stably surpassing naive OPD and RLVR baselines, thereby confirming that well-regulated signal quality, rather than mere teacher scale, governs successful exploration in OPD.

1 Introduction

On-policy distillation (OPD) has rapidly become a standard component of modern LLM post-training pipelines. By providing token-level supervision from a stronger teacher over the student's own rollouts, OPD has demonstrated remarkable effectiveness (Gu et al., 2026; Yang et al., 2025; DeepSeek-AI, 2026). Despite its growing adoption, the fundamental training dynamics of OPD remain poorly understood. Practitioners often observe highly inconsistent behaviors: OPD sometimes improves performance (Lu and Lab, 2025), yet in other cases becomes unstable (Luo et al., 2026; Zhu et al., 2026), collapses exploration, even underperforms outcome-based reinforcement learning (Li et al., 2026a). This raises our systematic empirical study, we aim to (1) explore what benefits OPD brings (Role§ 3), (2) when OPD fails (Pathologies§ 4), and (3) how to improve it from the previous findings (Regulations§ 5).

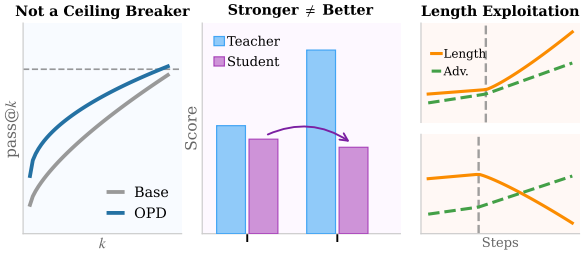

Figure 1: **Left**: OPD accelerates exploration but cannot exceed the model ceiling. **Middle**: a stronger teacher does not guarantee a better student. **Right**: advantage rises while length explodes or crashes.

Role OPD simultaneously embodies two distinct features: it is formulated as knowledge distillation (Gu et al., 2026), yet operates like an on-policy RL algorithm driven by dense token-level rewards (Yang et al., 2026a). This dual nature motivates us to investigate whether OPD acts as a standard distillation mechanism that expands the student's capability boundary (Hinton et al., 2015), or as an RL framework that merely reshapes trajectories within the base model's existing capability space (Chen et al., 2024; Yue et al., 2025). Our empirical evaluations strongly support the latter RL-centric view. We find that while OPD consistently accelerates early-stage learning, it rarely alters the asymptotic performance ceiling achievable by the base student. This acceleration stems from

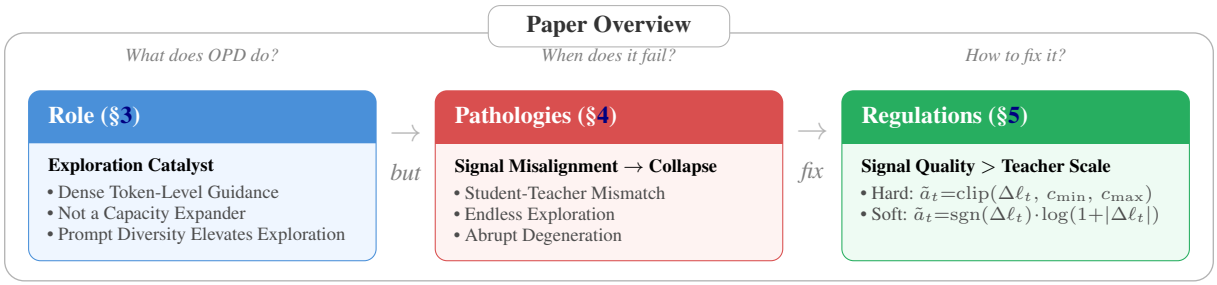

Figure 2: Overview. We characterize OPD's role as an exploration catalyst (§3), diagnose pathologies (§4), and propose lightweight signal regulations that restore stability without additional compute (§5).

the dense, token-level guidance provided along student-generated trajectories, which effectively helps the student discover correct reasoning paths already reachable within its inherent capability bounds. Furthermore, under a fixed computational budget, scaling problem diversity (prompt coverage) yields significantly higher accuracy than scaling per-problem sampling depth. Together, these findings demonstrate that OPD functions primarily as an exploration catalyst that reshapes the trajectory distribution, rather than a mechanism for fundamental capability expansion.

Pathologies Viewing OPD as an exploration guidance mechanism immediately uncovers a critical dependency: its effectiveness relies entirely on the fidelity of the guidance signal itself. While faithful teacher preferences steer the student toward correct trajectories, distorted signals inevitably bias exploration toward degenerate behaviors. Our analysis reveals two major sources of such guidance corruption, as shown in Figure 1.

- **Student-Teacher Mismatch**: When a severe capability gap exists between the teacher and the student, the teacher's preference allocations on student-generated trajectories decouple from actual quality, generating misleading guidance. This is rooted not in the teacher's absolute strength, but in the mismatch between teacher preferences and the student's rollout distribution.
- **Length Exploitation**: Because token-level signals are aggregated over sequences of varying lengths, the standard optimization objective inadvertently admits degenerate shortcuts. The student exploits this via either *Endless Exploration* (filler tokens to dilute negative advantages), or *Abrupt Degeneration* (premature truncation for high aggregated advantage).

Regulations We further explore how to eliminate these pathologies. Intuitive methods include cold-start teacher-student alignment finetuning (Luo et al., 2026; Li et al., 2026a), which are effective but require additional computation. We investigate two token-level regulation strategies designed to suppress pathological incentives while preserving useful exploration guidance: *Hard Clipping* and *Soft Log-scale Compression*. Across seven reasoning benchmarks, these interventions consistently stabilize training, eliminate previous pathologies, and substantially outperform naive OPD, RLVR, and prior OPD variants (Jin et al., 2026). Notably, our regulated framework with a modest teacher decisively surpasses state-of-the-art methods relying on a massive 30B teacher (Yang et al., 2026a; Hou et al., 2026), demonstrating that the quality of the optimization signal ultimately governs distillation success over brute-force teacher scaling. Overall, this work makes three contributions:

- We empirically frame OPD as an exploration guidance rather than a capability expander.
- We pinpoint guidance corruption, specifically distribution mismatch and length exploitation, as the root cause of OPD failures.
- We demonstrate that signal regulations effectively improve OPD stability and effectiveness across diverse reasoning tasks.

2 Preliminaries

2.1 On-Policy Distillation (OPD)

Let x denote a prompt and $\mathbf{y} = (y_1, \ldots, y_T)$ a response over vocabulary $\mathcal{V}$. We consider a fixed teacher π_{T} and a trainable student π_θ, with $\pi(\mathbf{y}|x) = \prod_{t=1}^{T} \pi(y_t|\mathbf{y}_{<t}, x)$.

Objective OPD trains the student to minimize the reverse KL divergence from student to teacher:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathrm{KL}[\pi_\theta \| \pi_{\mathrm{T}}] = -\mathbb{E}_{x,\, \mathbf{y}\sim\pi_\theta}\left[\log \frac{\pi_{\mathrm{T}}(\mathbf{y}|x)}{\pi_\theta(\mathbf{y}|x)}\right]. \tag{1}$$

Because the expectation is taken over the student's own samples, the signal is evaluated on sequences the student actually generates, avoiding the distributional mismatch inherent in off-policy distillation.

Per-token advantage Expanding the sequence-level KL token-by-token yields per-token log-ratio:

$$\Delta\ell_t \triangleq \log \frac{\pi_{\mathrm{T}}(y_t \mid \mathbf{y}_{<t}, x)}{\pi_\theta(y_t \mid \mathbf{y}_{<t}, x)}, \quad \mathbf{y} \sim \pi_\theta(\cdot \mid x). \tag{2}$$

Intuitively, $\Delta\ell_t > 0$ indicates the teacher assigns higher probability to token y_t than the student, providing a positive learning signal; $\Delta\ell_t < 0$ indicates the student over-weights y_t relative to the teacher.

Policy optimization OPD is implemented via policy gradient with $\Delta\ell_t$ as the per-token advantage. We apply clipping (Schulman et al., 2017) on the importance ratio to stabilize updates:

$$\mathcal{L}_{\mathrm{clip}}(\theta) = -\mathbb{E}_t[\min(r_t\,\tilde{a}_t,\ \mathrm{clip}(r_t, 1-\epsilon, 1+\epsilon)\,\tilde{a}_t)], \tag{3}$$

where $r_t = \pi_\theta(y_t \mid \mathbf{y}_{<t}, x) \,/\, \pi_{\theta_{\mathrm{old}}}(y_t \mid \mathbf{y}_{<t}, x)$ is the importance ratio and $\tilde{a}_t$ is a regulated form of $\Delta\ell_t$. This ratio clipping constrains how far the policy moves per update but does not alter the reward landscape defined by the teacher signal.

2.2 Empirical Framework and Setup

In this section, we establish the experiment bases for our study. To ensure a systematic investigation and the reproducibility of our findings, we strictly adhere to the experimental configurations described below across subsequent training and evaluation.

Training Configuration Our empirical analysis uses Qwen3-1.7B-Base as the student and Qwen3-series post-trained models of varying scales as teachers. To isolate true distillation gains from mere thinking-mode shifts, we enforce a strict non-thinking constraint by prefixing all rollouts with a block. This simultaneously ensures an aligned prompt format between models (Li et al., 2026a). Trained on the Nemotron-Cascade Math dataset (Wang et al., 2026), we compare models optimized via GRPO (Shao et al., 2024; Yu et al., 2025) and OPD (Gu et al., 2026).

Evaluation To provide a robust and statistically sound assessment, we employ the **avg@32** metric for AMC23 (MAA, 2026b), AIME 2024–2026 (MAA, 2026a), and HMMT25 Feb (Dekoninck et al., 2026), while reporting standard **pass@1** for MATH500 (Hendrycks et al., 2021) and Minerva (Lewkowycz et al., 2022) following previous works (Jin et al., 2026).

3 Exploring the Role of OPD

Conceptually, OPD operates at the intersection of knowledge distillation (Hinton et al., 2015; Gu et al., 2026), and on-policy RL with dense token-level rewards (Yang et al., 2026a). While standard distillation intuition suggests that the teacher transfers superior intelligence to expand the student's capability boundary, RL primarily reshapes trajectories within the latent capability space already encoded in the base model (Chen et al., 2024; Yue et al., 2025). Under large sampling budgets (k), the performance of RL-trained models often converges to that of their unaligned base counterparts via brute-force test-time sampling (Yue et al., 2025). This capability convergence motivates a fundamental question for OPD: *Does OPD fundamentally elevate the inherent capability ceiling of the student, akin to distillation, or does it accelerate early-stage exploration efficiency, much like RL?*

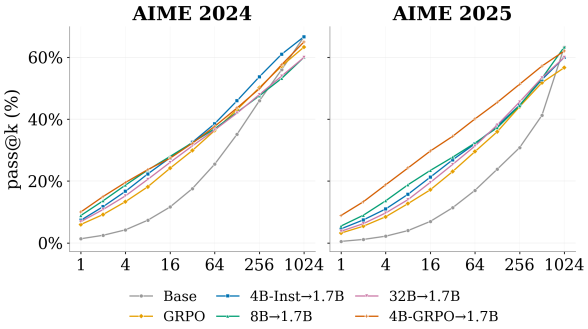

Figure 3: Pass@k performance of Qwen3-1.7B-Base trained via different methods. OPD variants outperform Base and GRPO models in low-k. However, the convergence at higher k suggests OPD accelerates exploration rather than raising the model's performance ceiling.

To disentangle these two possibilities, we systematically analyze the performance convergence trajectories of student models trained via OPD and RLVR across varying teacher configurations and evaluation sampling size k. We set the Qwen3-1.7B-Base as the student, and Qwen3-4B-instruct-2507, Qwen3-{4B,8B,32B} and GRPO trained Qwen3-4B as the teachers. Then we test the pass@1024 of tuned models to explore their performance ceilings.

Efficient Exploration rather than Ceiling Breaker As shown in Figure 3, OPD variants provide a distinct boost in the low-sample regime ($k \leq 64$). However, in the high-sample regime, this performance gap narrows. The base model eventually converges to the level of the GRPO and OPD-trained models, indicating that OPD, just like RLVR, does not fundamentally inject entirely new capabilities (Yue et al., 2025). Instead, it acts as an exploration catalyst that re-weights the student's in-

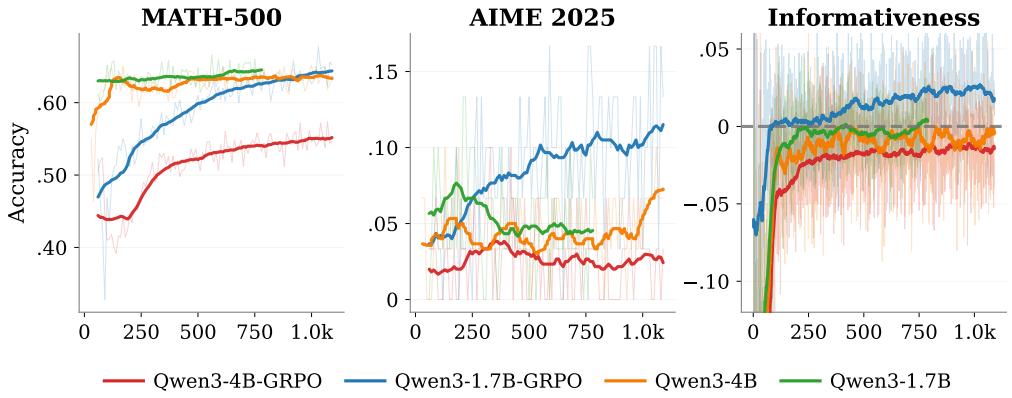

Figure 4: **Weak teacher outperforms strong teacher in OPD w/ Clip.** Benchmark accuracy (first two panels) and informativeness $\mathcal{I}$ on the validation set for OPD w/ Clip runs distilling from four teachers with varying capabilities into the same Qwen3-1.7B-Base student.

herent generative space, ensuring optimal solutions are surfaced within the first few attempts.

This characterization highlights a critical vulnerability: If the teacher's guidance is misaligned or the objective function contains structural flaws, this catalytic effect will instead accelerate the student's convergence toward degenerate or sub-optimal solutions. We dissect these specific pathologies in the following section.

Prompt Diversity v.s. Rollout Sizes Our previous analysis establishes that OPD benefits the student primarily through high-density guidance across diverse problem spaces. This motivates a critical resource-allocation question: Under a fixed computational budget, should we prioritize *problem coverage* (more distinct prompts) or *rollout depth* (more rollouts per prompt)?

To investigate this, we evaluate n=1 (B distinct prompts with 1 rollout each) against n=2 and n=8 configurations under identical total compute. As illustrated in Figure 5, the n=1 setup consistently achieves superior final accuracy. This confirms that OPD's exploration benefits scale primarily with problem diversity rather than per-problem sampling depth. Consequently, we adopt the $n = 1$ configuration across all subsequent experiments to maximize distillation efficiency.

Finding 1. OPD incentivizes the student's exploration efficiency, not performance ceilings.

4 Pathologies of On-Policy Distillation

Given that OPD operates by re-weighting the student's exploration landscape, its efficacy hinges on the *fidelity* of the reward signal. We identify two distinct failure modes where this guidance mechanism breaks down, leading to the stagnation or collapse of student performance.

First, a **Student-Teacher Mismatch** (§4.1): we challenge the stronger-is-better dogma by showing that a significant distributional gap between teacher and student can turn the distillation signal into counter-productive noise. Second, a **Length Exploitation** (§4.2): the cumulative nature of token-level advantages creates length-dependent shortcuts that allow the student to leverage the objective without genuine reasoning improvements.

4.1 Student-Teacher Mismatch

A common intuition in knowledge distillation is that a more capable teacher yields a superior student. However, our empirical results challenge this assumption, revealing a complex trade-off between standalone capability and distributional alignment.

Setup We continue to employ Qwen3-1.7B-Base as the student. To evaluate the student's sensitivity to varying teacher expertise, we select teachers across a controlled spectrum of capabilities (Table 1): Qwen3-{1.7B, 4B}-GRPO, and Qwen3-{1.7B, 4B}. The capability order of teachers is: Qwen3-4B-GRPO $>$ Qwen3-1.7B-GRPO $>$ Qwen3-4B $>$ Qwen3-1.7B.

Early Efficiency is Deceptive As shown in the MATH-500 results (Figure 4), weaker teachers such as Qwen3-1.7B and Qwen3-4B lead to rapid accuracy gains during the initial training steps. This high initial learning efficiency suggests that teachers with lower performance are more learnable for a base student. However, this early speed is deceptive. These models quickly hit a performance plateau, particularly on challenging benchmarks, due to their capability boundaries. This observation

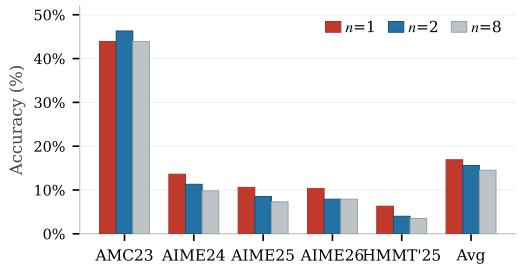

Figure 5: **Avg@32 across varying prompt-to-rollout configurations.** Under a fixed batch B, we vary the rollout count n per prompt. Maximizing problem diversity consistently outperforms a larger sampling size.

implies that initial learning velocity is an unreliable proxy for the final capability ceiling.

The Failure of the Strongest Teacher The most striking observation is the failure of the most capable model in our suite: Qwen3-4B-GRPO. Despite having the highest performance, it is paradoxically the least effective teacher. As shown across both benchmarks in Figure 4, the student distilled from 4B-GRPO remains nearly stagnant, with its accuracy on AIME 2025 hovering near 2% throughout the training process.

This inversion is a significant anomaly: a 4B RL-trained model provides significantly worse guidance than its 1.7B counterpart. It suggests that standalone performance does not translate into teaching quality. Ultimately, 1.7B-GRPO yields the highest final accuracy, outperforming teachers that are either easier to learn from or smarter in isolation (4B-GRPO). This indicates that the most effective teacher is not necessarily the one who is easiest to mimic or the one with the strongest performance, but the one whose reasoning paths are both advanced and bridgeable for the student.

Diagnosis To explain this paradox, we introduce the **Informativeness** metric $\mathcal{I}$, which measures whether the teacher's token-level signal discriminates correct from incorrect student rollouts:

$$\mathcal{I} = \mathbb{E}_{\mathbf{y}\sim\pi_\theta}\big[\,\overline{\Delta\ell} \mid r=1\big] - \mathbb{E}_{\mathbf{y}\sim\pi_\theta}\big[\,\overline{\Delta\ell} \mid r=0\big], \quad (4)$$

where $\overline{\Delta\ell} = \frac{1}{|\mathbf{y}|}\sum_t(\log\pi_T(y_t) - \log\pi_\theta(y_t))$ is the mean per-token log-ratio, r denoting the rollout accuracy. A positive $\mathcal{I}$ indicates that the teacher assigns a higher probability to tokens in correct rollouts than incorrect ones, providing an outcome accuracy aligned optimization. The rightmost panel of Figure 4 reveals that informativeness ($\mathcal{I}$) precisely predicts distillation trajectories, clustering into three distinct dynamic modes: **(1) Passable** ($\mathcal{I} \approx 0$)**:** Qwen3-1.7B and 4B drive rapid early gains on MATH500 but quickly plateau on the harder AIME 2025; **(2) Misleading** ($\mathcal{I} < 0$)**:** Qwen3-4B-GRPO provides persistently anti-correlated signals, leading to early accuracy drops and a low ceiling; and **(3) Recovery**: Qwen3-1.7B-GRPO starts negative but recovers past zero, unlocking a late-stage surge that ultimately outperforms all others.

Finding 2. The effectiveness of OPD depends on whether the teacher can give rewards aligned with Informativeness $\mathcal{I}$ on student's rollouts, rather than the teacher's absolute capability.

4.2 Length Exploitation

While the previous section highlighted how external teacher selection destabilizes training, we now turn to an intrinsic vulnerability embedded within the standard OPD objective itself. The sequence-level guidance is propagated via the token-level distillation advantage, $a_t = \log\pi_T(y_t|y_{<t}) - \log\pi_\theta(y_t|y_{<t})$. Unlike outcome-based rewards that are inherently bounded and strictly tied to final correctness, a_t is unconstrained, exhibits high variance, and is fundamentally decoupled from the actual accuracy of the response. Since the token-mean advantage $\bar{a} = \frac{1}{T}\sum_t a_t$ depends on both the per-token signal and the response length T, the student can exploit this interaction to inflate its advantage without improving accuracy. We identify two complementary failure modes, illustrated in Figure 6 and Figure 7.

Endless Exploration (Mode A) When a rollout yields an incorrect reasoning path, it incurs a severe overall negative advantage. Crucially, because standard OPD objectives optimize the sequence-averaged advantage $\bar{a} = \frac{1}{T}\sum_{t=1}^{T} a_t$, the student discovers a degenerate shortcut to evade this penalty: appending low-information filler tokens to inflate the denominator T. Formally, suppose the core reasoning block terminates at step T_0 with a cumulative negative signal $A_{\text{core}} = \sum_{t=1}^{T_0} a_t < 0$. By continually generating redundant tokens up to a prolonged length T ($T \gg T_0$), the aggregated advantage decomposes as:

$$\bar{a} = \frac{A_{\text{core}}}{T} + \frac{T-T_0}{T}\,\bar{a}_{\text{filler}}. \quad (5)$$

Eq. (5) reveals a severe structural pathology: By scaling the sequence length T, the student can asymptotically suppress the magnitude of its negative advantage ($\lim_{T\to\infty}\bar{a} = 0$), effectively wash-

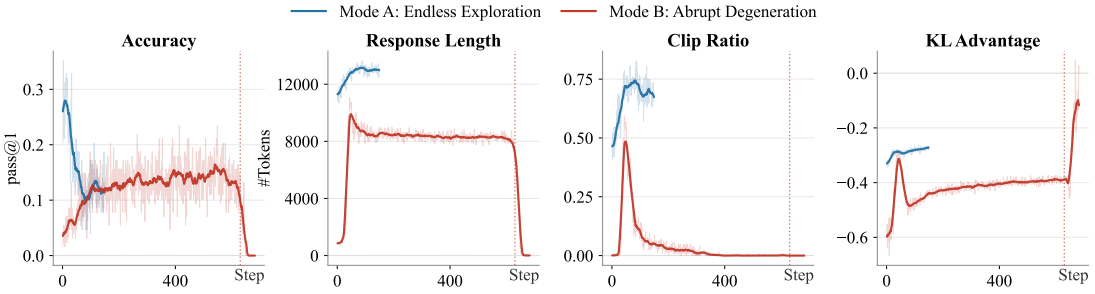

Figure 6: Length exploitation illustration. Advantage increases, while accuracy drops. Teacher→Student: Qwen3-4B-GRPO→Qwen3-1.7B (Mode A), and Qwen3-1.7B-GRPO→Qwen3-1.7B-Base (Mode B).

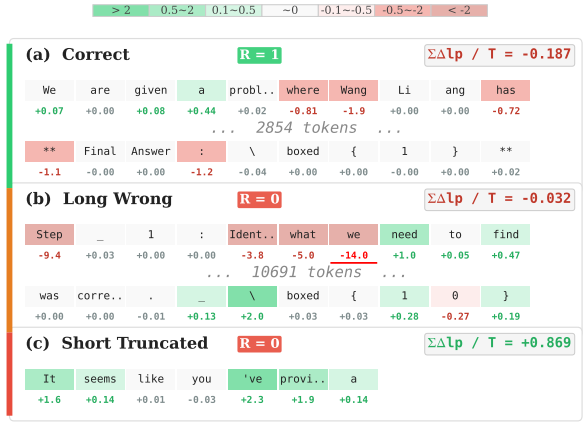

Figure 7: **(a)** A correct rollout receives a mildly negative token-mean. **(b)** *Mode A (Endless Exploration)*: a long wrong rollout with negative scores diluted across tokens, failing to penalize the incorrect response. **(c)** *Mode B (Abrupt Degeneration)*: a short, fully teacher-preferred rollout prefix that yields a strongly positive token-mean despite an incorrect outcome.

ing out the penalty gradient ($\nabla_\theta \mathcal{L} \approx 0$). As a result, the student is never penalized for its logical failures, trapping the training loop in an unproductive state of endless, verbose exploration without any actual improvement in reasoning accuracy.

As shown in Figure 6, both runs exhibit a common pattern: response lengths grow sharply. Once truncated, the model fails to produce valid final answers, causing a sharp collapse in accuracy. This pathology is particularly severe in the Qwen3-4B→Qwen3-1.7B setting, which response length grows until hitting the predefined maximum limit. Crucially, despite the drop in performance, the aggregated advantage continues to climb, demonstrating a clear pattern of advantage hacking.

Short-Sequence Exploitation (Mode B) In this situation, the student learns to truncate responses prematurely. By generating a high-confidence preamble (e.g., "We are given the problem...") and immediately emitting `EOS`, the model captures the favorable prefix advantage $\sum_{t=1}^{T_{\text{pre}}} a_t$ while avoiding the riskier reasoning steps where a_t may turn negative. Formally, whenever the prefix advantage exceeds the full-sequence advantage:

$$\frac{1}{T_{\text{pre}}} \sum_{t=1}^{T_{\text{pre}}} a_t > \frac{1}{T} \sum_{t=1}^{T} a_t, \tag{6}$$

The optimizer is incentivized to favor shorter trajectories regardless of correctness. As shown in Figure 6, in the late training stage, the student only output extremely short and safe responses, but receives a high reverse KL advantage from the teacher. Once the policy locks into short outputs, the gradient reinforces this equilibrium.

Dignosis Both modes exploit the same structural flaw: the token-level advantage a_t interacts with sequence length in a way that creates *gradients unrelated to reasoning correctness*. Mode A exploits unfavorable sequences by padding late; Mode B exploits favorable prefixes by cutting early. Together they form a squeeze: the optimizer can always improve the objective by manipulating length rather than improving reasoning, regardless of whether the current rollout outcome is positive or negative.

Finding 3. The unstable token-level objective introduces degenerate optimization shortcuts at both length extremes.

5 Regulations

Prior work stabilizes OPD through mechanisms such as utilizing off-policy data (Luo et al., 2026), or scoring top-k tokens beyond the predicted one to smooth gradient estimates (DeepSeek-AI, 2026). These approaches are effective but incur substantial overhead, maintaining full vocab replay buffers during training (DeepSeek-AI, 2026) or warm-starting students with SFT on teacher rollouts for cold

Teacher	Student	Method	*pass@1*		*avg@32*					Avg
			MATH500	Minerva	AMC'23	AIME'24	AIME'25	AIME'26	HMMT'25	
—	Qwen3-1.7B-Base	–	18.0	4.1	6.3	1.7	0.4	0.6	0.1	4.5
	Qwen3-1.7B	–	73.0	32.7	42.5	13.4	9.8	7.8	3.7	26.1
	Qwen3-8B	–	83.4	47.8	66.9	29.1	20.9	14.8	10.4	39.0
	Qwen3-4B-Inst-2507	–	93.6	52.9	94.2	59.3	47.4	55.8	31.4	62.1
—	Qwen3-1.7B-Base	GRPO	63.1	29.5	43.0	9.2	4.9	4.7	2.3	22.4
—	Qwen3-1.7B	GRPO	90.0	45.9	81.7	41.9	32.5	37.5	21.7	50.2
—	Qwen3-4B	GRPO	93.6	55.5	92.9	64.4	57.1	57.3	37.3	65.4
Teacher: Qwen3-8B										
	Qwen3-1.7B-Base	KD	63.8	30.2	40.6	9.2	4.6	–	–	–
	Qwen3-1.7B-Base	EOPD	68.7	30.2	41.9	10.4	5.8	–	–	–
Teacher: Qwen3-1.7B-GRPO										
	Qwen3-1.7B-Base	OPD	75.6	36.3	43.9	13.6	10.6	10.3	6.3	28.1
	Qwen3-1.7B-Base	+Clip	77.0	35.4	45.7	12.1	13.7	10.5	5.8	28.6
	Qwen3-1.7B-Base	+Scale	**79.3**	**40.1**	**46.3**	**13.8**	**15.6**	**10.6**	**7.3**	**30.4**
Teacher: Qwen3-4B-GRPO										
	Qwen3-1.7B-Base	OPD	74.4	**34.2**	44.3	11.6	9.4	8.6	4.7	26.7
	Qwen3-1.7B-Base	+Clip	**76.2**	33.8	**47.4**	10.8	**13.5**	**8.7**	**6.5**	**28.1**
	Qwen3-1.7B-Base	+Scale	74.3	**34.2**	41.2	**12.8**	12.1	7.8	5.4	26.8

Table 1: Performance of different methods on benchmarks. Among all evaluated downstream student models (bottom two blocks), the overall best scores are **bolded** and the second best are underlined.

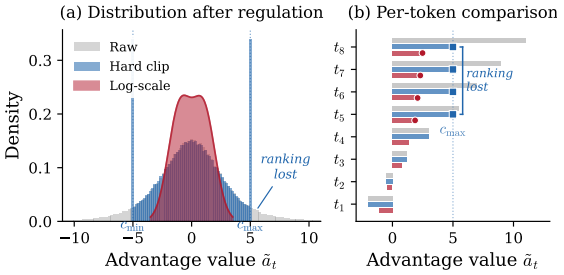

Figure 8: Hard Clipping collapses all tokens beyond the $[c_{\min}, c_{\max}]$ to the same value (ranking lost). Log-scale compression preserves ordering across the full range.

start (Luo et al., 2026). We instead pursue modifications that stabilize OPD *without additional compute*, operating entirely within the existing training loop and motivated by the diagnostic findings of the previous section.

5.1 Stabilizing OPD with Signal Regulation

Motivated by the aforementioned pathologies, we transition from diagnosing length exploitation to developing proactive in-loop remedies. To prevent extreme token-level log-ratios from dominating the policy gradient, which **exacerbates the teacher–student mismatch and enables the student to exploit length through extreme signals**, we introduce token-level signal regulation.

Our design follows two principles: **(1) Preference preservation**, regulation must not reverse the teacher's ordinal signal; **(2) Outlier suppression**, pathologically large magnitudes should be dampened without discarding fine-grained token-level information. We investigate two complementary paradigms, representing *Hard truncation* and *Soft compression* respectively (Figure 8):

- **Hard Clipping.** As a representative of rigid thresholding, this method truncates the advantage into a deterministic bound to eliminate severe outliers while preserving signal diversity:

$$\tilde{\Delta}\ell_t = \mathrm{clip}\big(\Delta\ell_t,\ c_{\min},\ c_{\max}\big). \tag{7}$$

- **Soft Log-Scale Compression.** As a smooth alternative, this method dampens unbounded signals without discarding the relative rankings beyond fixed bounds:

$$\tilde{\Delta}\ell_t = \mathrm{sign}(\Delta\ell_t) \cdot \log\big(1 + |\Delta\ell_t|\big). \tag{8}$$

The choice of $\log(1+|x|)$ is motivated by its dual-regime properties. Near zero, it approximates a linear mapping ($\tilde{\Delta}\ell_t \approx \Delta\ell_t$), fully preserving the teacher's fine-grained preferences. Crucially, when $|\Delta\ell_t| \to \infty$, its sub-linear growth radically compresses the magnitude of unbounded values.

Experiment Setup To explore the performance of the above two optimization techniques for OPD, we utilize Qwen3-1.7B-GRPO, Qwen3-4B-GRPO

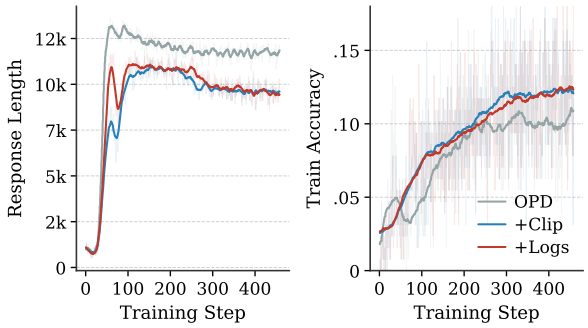

Figure 9: Training dynamics of OPD variants. Teacher: Qwen3-4B-GRPO; Student: Qwen3-1.7B-Base. Vanilla OPD suffers greater length exploitation, whereas Clip produces **shorter responses and achieves higher accuracy** than naive OPD.

as the teachers, and Qwen3-1.7B-Base and Qwen3-1.7B as the students. For better comparison, we select current OPD variants as counterparts (Jin et al., 2026; Yang et al., 2026a; Hou et al., 2026). The performance of Qwen3-1.7B-Base as the student is shown in Table 1, and the scores of Qwen3-1.7B are shown in Table 2.

5.2 Analysis

Signal Regulation is Effective. As shown in Table 1, both regulation variants yield clear improvements over naive OPD. With the 1.7B-GRPO teacher, OPD+Clip brings substantial gains across all benchmarks. Comparing the two variants, log-scale compression outperforms hard clipping for Qwen3-1.7B-Base, likely because it retains the teacher's full token ranking rather than flattening extreme values to the clip boundary.

Alleviating the Student-Teacher Mismatch. Our experiments further demonstrate that naively scaling up the teacher does not guarantee better student performance. Hard clipping partially rescues this, lifting the 4B-guided student back to 28.1 of average score. Interestingly, log-scale compression, which excelled with the 1.7B teacher, fails to yield comparable gains here, as the training dynamics shown in Figure 9. We attribute this to the nature of the distributional shift: a much stronger teacher heavily mixes genuine correctness cues with biases in its advantages. Log-scale compression preserves this noise, whereas hard clipping aggressively truncates it. This yields a crucial insight: **when the teacher-student gap is large, aggressively truncating extreme signals is more effective than compressing them.**

Method	AIME24	AIME25	AIME26	HMMT25
Teacher: Qwen3-30B-A3B-Instruct-2507				
Teacher	74.7	62.8	63.3	44.2
UniOPD	35.2	30.7	–	17.7
GOPD	37.3	31.5	–	16.2
Teacher: Qwen3-1.7B-GRPO				
Teacher	41.9	32.5	37.5	21.7
CLIP	<u>44.4</u>	<u>32.5</u>	<u>35.9</u>	**20.9**
Teacher: Qwen3-4B-GRPO				
Teacher	64.4	57.1	57.3	37.3
CLIP	**45.2**	**37.5**	**36.0**	<u>19.0</u>

Table 2: Performance of Qwen3-1.7B student tuned by OPD+Clip (CLIP) and previous methods (Hou et al., 2026; Yang et al., 2026a), which further proves that *a stronger teacher cannot ensure better student performance*. The **best** and the <u>second best</u> scores are marked.

Impact of Student Capacity Comparing Tables 1 and 2 reveals a striking impact between student capacity and teacher strength. When the student is upgraded from a base model to the post-trained Qwen3-1.7B, the initial distributional mismatch is significantly mitigated. Consequently, the performance hierarchy flips: Qwen3-4B-GRPO teacher now delivers the best results, decisively outperforming the Qwen3-1.7B-GRPO teacher. This shift confirms that the trap stems from the capability gap with student-teacher mismatch. Crucially, by resolving this mismatch, our regulated OPD with a 4B teacher ultimately surpasses previous methods distilled from a larger 30B teacher. This demonstrates that proper capacity matching combined with signal regulation can eclipse brute-force parameter scaling.

6 Related Work

On-Policy Distillation. Early formulations of OPD optimized a reverse-KL objective to restrict the student to the teacher's high-probability modes (Gu et al., 2026; Agarwal et al., 2024). Subsequent work reframed the teacher's per-token log-ratios as dense rewards capable of shaping exploration trajectories (Yang et al., 2026a), fueling its rapid adoption in scale post-training (DeepSeek-AI, 2026; Yang et al., 2025; GLM-5-Team et al., 2026; Yang et al., 2026b). Despite this widespread use, the failure conditions of this dense guidance remain poorly understood. Prior exploratory studies attribute training instability to factors like thinking-pattern compatibility, offering external fixes such as prompt alignment or cold-start initialization (Li

et al., 2026a; Luo et al., 2026; Yu et al., 2026). Our work takes a fundamentally parallel yet deeper trajectory: we diagnose signal-level pathologies (student-teacher mismatch and length exploitation) and propose zero-overhead in-loop regulations (clipping and log-scale compression) that require neither off-policy data nor prompt filtering.

Capacity Gap in Knowledge Distillation. A recurring phenomenon in knowledge distillation (KD) is that excessively capable teachers can unexpectedly degrade student performance. Specifically, Cho and Hariharan (2019) noted that higher benchmark accuracy in teachers does not inherently translate into superior distillation efficacy. To mitigate the structural mismatches between models, various strategies have been proposed: Huang et al. (2022) identified a severe prediction discrepancy between teachers and students, proposing the utilization of the teacher's recommendation label preference order to alleviate the gap, while Mirzadeh et al. (2019) introduced intermediate teacher assistants to bridge this capacity rift. Within the context of reasoning, this learnability barrier remains equally pronounced. For instance, Li et al. (2025) observed that complex, long reasoning traces generated by strong teachers consistently underperform simpler, teacher-free approaches when distilled into smaller student models. Li et al. (2026b) demonstrated that the alignment of thinking patterns between teachers and students profoundly impacts distillation outcomes.

Our work extends this classic line of inquiry into the on-policy regime. Rather than merely documenting the empirical decay caused by strong teachers, we offer a precise, signal-level explanation for *why* and *when* these advanced guidance signals corrupt on student-generated rollouts. Furthermore, we demonstrate that the capability gap can be mitigated via lightweight, in-loop signal regulation.

7 Conclusion

In this work, we look inside OPD to shift the paradigm from empirical scaling to rigorous signal-level diagnosis. We first established that OPD acts as an *exploration catalyst* which benefits more from problem diversity. Then we systematically expose two critical pathologies: *length exploitation* via redundant token padding, and the *Student-Teacher Mismatch*, where severe capabilities mismatch causes negative impacts. To restore guidance fidelity, we introduce lightweight signal regulations that require zero off-policy overhead. Our empirical results demonstrate that regulating token-level advantage dynamics effectively eliminates these failures, enabling a 4B teacher to surpass configurations distilled from larger models, which is *signal quality, not teacher scale, governs OPD success*.

Limitations

Although our proposed hard clipping and soft log-scale compression strategies effectively restore guidance fidelity and stabilize training, they introduce hyperparameter heuristics that require empirical tuning based on the specific teacher-student capacity gap. A dynamic, adaptive regulation framework could be further explored. Our systematic empirical study primarily focuses on mathematical reasoning benchmarks, where outcomes are easily verifiable and reasoning chains are dense. The behavior of regulated OPD in open-ended text generation or knowledge-intensive QA tasks has not been fully verified.

References

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. GKD: Generalized knowledge distillation for auto-regressive sequence models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Qiguang Chen, Libo Qin, Jiaqi WANG, Jingxuan Zhou, and Wanxiang Che. 2024. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4794–4802.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. 2026. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, and 167 others. 2026. Glm-5:

from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2026. Minillm: On-policy distillation of large language models. *Preprint*, arXiv:2306.08543.

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

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

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, and 1 others. 2026. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. *Preprint*, arXiv:2205.10536.

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

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *NeurIPS*.

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

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria. Association for Computational Linguistics.

Zhaoyi Li, Xiangyu Xi, Zhengyu Chen, Wei Wang, Gangwei Jiang, Ranran Shen, Linqi Song, Ying Wei, and Defu Lian. 2026b. On the role of reasoning patterns in the generalization discrepancy of long chain-of-thought supervised fine-tuning. *Preprint*, arXiv:2604.01702.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/on-policy-distillation.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. 2026. Demystifying opd: Length inflation and stabilization strategies for large language models. *Preprint*, arXiv:2604.08527.

MAA. 2026a. American invitational mathematics examination (AIME). https://www.maa.org/math-competitions.

MAA. 2026b. American mathematics competitions (AMC). https://www.maa.org/math-competitions.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2019. Improved knowledge distillation via teacher assistant. *Preprint*, arXiv:1902.03393.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yang Chen, Yangyi Chen, Zhuolin Yang, Zihan Liu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2026. Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models. *Preprint*, arXiv:2512.13607.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026a. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *Preprint*, arXiv:2602.12125.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, Renjie Pi, Grace Lam, Nayeon Lee, Alexander Bukharin, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2026b. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and

16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, Qunzhong Wang, Yang Shi, Xiaobin Hu, Xiangyu Yue, Jiaqi Wang, and Shuicheng Yan. 2026. Dopd: Dual on-policy distillation. *Preprint*, arXiv:2606.30626.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. 2026. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *Preprint*, arXiv:2605.11182.

A Experiment Configurations

We implement the training pipeline using verl. For the OPD pipeline, policy optimization uses a total batch size of 128, 1 rollout for each prompt. For the GRPO baselines, the group advantage is calculated across a size of 128×8 rollouts. Optimization is limited to a single epoch per rollout batch to eliminate off-policy bias. The maximum sequence length is set to $16,384$ tokens during training. Rollouts are generated using stochastic sampling with a temperature of $\tau = 1.0$ and top-$p = 1.0$.

For the avg@32 metric, we utilized $N = 40$ candidate responses per problem for better estimation. To obtain a stable estimation, we perform $M = 10$ independent trials. In each trial, we randomly sample a subset of $k = 32$ responses from the candidate pool and calculate the correctness of the j-th response as $\mathbb{I}_j \in \{0, 1\}$. The final score is the average across all trials, formulated as:

$$\text{avg@32} = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{k} \sum_{j \in \mathcal{S}_i} \mathbb{I}_j \right) \tag{9}$$