BEYOND SELF-EXPLORATION: STABLE OFF-POLICY REINFORCEMENT LEARNING FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become a central paradigm for aligning LLMs, yet most existing approaches rely on on-policy self-exploration, which can lead to optimization stagnation when the policy is weak due to low-quality rollouts. Leveraging high-quality external trajectories offers a natural remedy, but this off-policy setting is notoriously unstable. Through an empirical analysis of PPO and GRPO, we identify *unregulated importance ratios* induced by off-policy distributions as a primary cause of this instability. Beyond instability, off-policy RL also suffers from *attenuated gradients*, as informative off-policy tokens typically receive low probability under the target policy. To address these challenges, we introduce **Adaptive Projected Robust Optimization (APRO)** to enable stable off-policy RL in LLM optimization. APRO constructs a tractable proxy for the unknown off-policy distribution and adaptively switches to a local policy anchor when proxy importance-ratio estimates become unreliable. In addition, APRO introduces discrepancy-aware advantage scaling to amplify informative learning signals that would otherwise be suppressed. We evaluate APRO on mathematical reasoning and instruction-following tasks in fully off-policy settings. Experimental results show that APRO achieves stable optimization and improves over the evaluated offline distillation and ratio-based off-policy RL baselines. Further experiments show that APRO can be used in mixed-policy training and can provide stronger initializations for subsequent on-policy RL, highlighting stable off-policy RL as a useful capability-transfer stage for weak target policies.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a dominant paradigm for post-training large language models (LLMs). In practice, this post-training stage is predominantly implemented using clipped policy optimization methods, with objectives derived from Proximal Policy Optimization (PPO) (Schulman et al., 2017) and its variants serving as the standard backbone (Ouyang et al., 2022; DeepSeek-AI, 2025). These methods have led to substantial progress in mathematical reasoning and instruction following. However, most existing RL pipelines for LLMs remain predominantly *on-policy*: trajectories are sampled from the current policy or a recent snapshot, and policy updates are stabilized using importance ratios with respect to this known sampling policy. This on-policy assumption becomes restrictive when the target model is weak. In challenging reasoning tasks, a small policy may rarely sample correct solutions, making reward signals sparse and self-exploration inefficient. As a result, the model can reinforce behaviors that already lie within its own distribution, but struggles to acquire reasoning strategies that it almost never discovers by itself. Recent studies on off-policy and hybrid-policy LLM training similarly observe that relying exclusively on self-generated rollouts can limit the acquisition of new behaviors, while external trajectories from stronger policies can provide useful guidance beyond the model's initial capability boundary (Yan et al., 2025).

A natural way to overcome this bottleneck is to decouple trajectory generation from policy optimization. Instead of requiring the target model to discover useful trajectories through self-exploration, one can train from trajectories generated by stronger external policies. This gives rise to a practically important setting: *black-box fully off-policy RL*. The learner is given a static dataset of prompt-response trajectories and rewards, but does not have access to the behavior policy's logits or sampling probabilities. This setting arises naturally when trajectories are collected from proprietary 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.

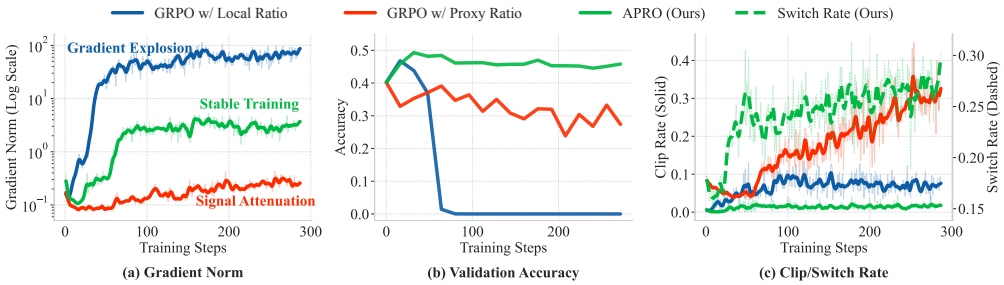

Figure 1: **(a) Gradient Norm:** The local-anchor baseline (blue) exhibits rapid gradient growth. Conversely, a standard proxy-ratio off-policy objective (red) exhibits *signal attenuation*. APRO (green) maintains stable gradient flow. **(b) Validation Accuracy:** These instabilities result in catastrophic collapse (blue) or unstable convergence (red), whereas APRO (green) achieves stable improvement. **(c) Clip/Switch Rate**[1]**:** The proxy-ratio baseline (red) exhibits a high and rising clip rate, indicating frequent trust-region violations. APRO (green) actively switches in unsafe regimes (green dashed line) and maintains a low effective clip rate comparable to the local trust-region anchor.

models, public model outputs, human or expert demonstrations, or historical rollout logs where behavior probabilities were not stored.

A straightforward solution is to apply supervised fine-tuning (SFT) or other offline distillation-style objectives to these external trajectories. Such methods provide strong cold-start signals and are important baselines in our experiments. They increase the likelihood of selected trajectories, possibly with reward or advantage weights, but do not explicitly correct the discrepancy between the target policy and the behavior policy. Value-based methods such as TBRM (Yuan et al., 2025) offer another route, using a trajectory-level Bellman residual without behavior importance ratios. Our question is how to stabilize PPO/GRPO-style policy-gradient updates on black-box external trajectories, and whether this route provides benefits over these alternatives.

Extending policy-gradient methods to this setting is challenging. *First*, the behavior distribution is unknown, making standard importance sampling corrections intractable. If the distribution shift is ignored and external trajectories are treated as on-policy samples, the resulting estimator becomes severely biased and may lead to gradient explosion and performance collapse, as shown in Figure 1(a,b). *Second*, approximating the behavior distribution with a projected proxy policy only partially resolves the issue. The proxy can be inaccurate in low-support regions, and the resulting importance ratios may frequently leave the trust region, producing unstable updates under severe distribution mismatch (Fujimoto et al., 2019). *Third*, even when the ratio is controlled, off-policy learning suffers from *signal attenuation*. The most informative decisions in external trajectories often correspond to tokens that the target model assigns very low probability. Practical clipping and variance-control mechanisms can then suppress precisely the low-probability but high-reward updates that are needed for capability transfer (Wang et al., 2025).

In this paper, we propose **Adaptive Projected Robust Optimization** (APRO), a stable RL method for learning from black-box external trajectories. APRO addresses the above challenges through three complementary mechanisms. First, *projected behavior estimation* trains a proxy policy on the external trajectories, providing a tractable denominator. Second, *adaptive ratio switching* combines proxy-based correction with a local policy anchor when importance-ratio estimates become unreliable. Third, *discrepancy-aware advantage scaling* amplifies positive updates on high-discrepancy trajectories and reduces penalties on negative trajectories that the model already assigns relatively low probability. These components determine the denominator, select the ratio, and weight its learning signal, respectively.

We evaluate APRO on mathematical reasoning and instruction-following tasks under black-box off-policy training. The experiments compare against regression-style offline objectives, including RWR (Peters & Schaal, 2007), AWR (Peng et al., 2019), and NLL-UL (Welleck et al., 2019), as well as recent off-policy RL methods such as TOPR (Roux et al., 2025), ASPO (Wang et al.,

[1]The *clip rate* and *switch rate* denote the proportions of importance ratios that trigger the clipping mechanism and our proposed adaptive ratio switching mechanism, respectively.

2025), DCPO (Yang et al., 2025), and BAPO (Xi et al., 2025). APRO improves over imitation-style objectives and several black-box off-policy policy-gradient baselines, suggesting that its gains do not merely come from fitting external trajectories. The results also indicate that fully off-policy RL is not meant to replace on-policy self-exploration. Rather, it is most useful as a capability-transfer stage for weak policies whose own rollouts are initially uninformative.

Additional continued-RL and mixed-policy experiments support this interpretation. APRO-trained checkpoints can be further improved by on-policy GRPO, and external trajectories remain useful even when mixed with on-policy rollouts. These results suggest that stable off-policy policy optimization can transfer useful external behaviors into the target policy and prepare it for more effective subsequent reinforcement learning.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

We consider post-training a large language model (LLM) with parameters θ to maximize a scalar reward function $R(q, o)$ defined over prompt–response pairs. The model is viewed as a policy π_θ that, given a prompt q, generates a response $o = (o_1, \ldots, o_T)$ autoregressively as $\pi_\theta(o|q) = \prod_{t=1}^{T} \pi_\theta(o_t|c_t)$, where $c_t = (q, o_{<t})$ denotes the generation context at step t. The standard reinforcement learning objective is to maximize the expected reward: $\mathcal{J}(\theta) = \mathbb{E}_{q,o\sim\pi_\theta}[R(q, o)]$. Policy gradient methods typically optimize this objective through iterative updates. At iteration k, trajectories are sampled from the previous policy $\pi_{\theta_{k-1}}$ (or $\pi_{\theta_{\text{old}}}$), and the policy is updated to π_{θ_k}.

2.2 PROXIMAL POLICY OPTIMIZATION (PPO)

Proximal Policy Optimization (PPO) (Schulman et al., 2017) is a widely used policy gradient method that stabilizes training through a clipped surrogate objective. It estimates policy gradients using importance sampling over trajectories from $\pi_{\theta_{\text{old}}}$, with the importance ratio at time step t defined as $r_t = \frac{\pi_\theta(o_t|c_t)}{\pi_{\theta_{\text{old}}}(o_t|c_t)}$. The PPO objective maximizes a lower bound on policy improvement:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},o\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{T}\sum_{t=1}^{T}\min\Big(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t\Big)\right], \tag{1}$$

where A_t denotes an advantage estimate (e.g., computed using Generalized Advantage Estimation (GAE)), which may leverage a learned value function $V_\psi(c_t)$ to balance bias and variance. PPO effectively constrains policy updates within an approximate trust region controlled by ϵ.

2.3 GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) adapts PPO for LLM post-training by eliminating the explicit value function and instead estimating advantages via relative comparisons among multiple samples for the same prompt. Specifically, for each prompt q, GRPO samples a group of G responses $\{o_i\}_{i=1}^{G}$ from $\pi_{\theta_{\text{old}}}$ and computes the advantage for the i-th response using group statistics: $\hat{A}_i = \frac{R(q,o_i)-\text{mean}(\{R_j\})}{\text{std}(\{R_j\})}$. The resulting objective mirrors PPO's clipped formulation:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_i, \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big)\right], \tag{2}$$

where the advantage $\hat{A}_i$ is shared across all tokens in the sequence o_i. A KL divergence penalty is often included to regularize the learned policy toward a reference model π_{ref}.

2.4 OFF-POLICY OPTIMIZATION: CHALLENGES

We now consider optimizing a policy from a static off-policy dataset $\mathcal{D}_{\text{off}}$ generated by an external behavior distribution $\mu(o|q)$. If μ were known, a natural off-policy variant of PPO/GRPO would replace the on-policy denominator $\pi_{\theta_{\text{old}}}$ with μ, yielding the ratio $r_t^\mu(\theta) = \pi_\theta(o_t|c_t)/\mu(o_t|c_t)$. In the black-box setting, however, the behavior probabilities are unavailable. Treating the external trajectories as on-policy samples gives a stable local ratio, but it no longer corrects the mismatch between the target policy and the true data-generating distribution. Conversely, estimating μ with

a proxy policy provides a behavior-aware denominator, but the resulting ratio can be unreliable in low-support regions, where small denominator errors lead to large or over-clipped updates.

A second difficulty is that the most useful external actions are often those assigned low probability by the weak target policy. These tokens may correspond to crucial reasoning steps that the target model rarely discovers through self-exploration. Standard clipped objectives tend to suppress such large off-policy deviations, causing the learning signal of informative positive samples to be attenuated. At the same time, uniformly penalizing negative off-policy tokens can waste update budget on actions that the target policy already considers unlikely. Therefore, off-policy RL requires an objective that both regulates unreliable importance ratios and preserves useful learning signals.

3 METHODOLOGY

This work centers on the problem of optimizing a target policy π_θ via off-policy reinforcement learning using a static dataset $\mathcal{D}_{\text{off}}$. The dataset is drawn from a high-quality *behavior distribution* μ generated by external sources, such as strong proprietary models, human experts, or heterogeneous solver mixtures, whose exact probability densities may be unknown. Formally, the dataset consists of prompts paired with their corresponding off-policy responses: $\mathcal{D}_{\text{off}} = \{(q_i, \{o_{i,j}\}_{j=1}^{T})\}_{i=1}^{N}$, where q_i denotes a query and $\{o_{i,j}\}$ are trajectories generated by $\mu(\cdot \mid q_i)$.

This setting gives rise to **three** fundamental challenges. (1) *Intractability of importance weights*: the agnostic nature of μ may prevent direct computation of standard importance sampling ratios. (2) *Unregulated importance ratios*: the potentially large distributional shift between the weak target policy π_θ and the strong behavior distribution μ leads to high-variance importance ratios and gradient estimates. (3) *Signal attenuation*: informative tokens from external trajectories may be assigned vanishingly small probability under the target π_θ, leading to negligible learning signal.

To address these challenges, we propose **Adaptive Projected Robust Optimization (APRO)**, a unified framework that combines *projected behavior estimation* for tractable likelihood approximation, *adaptive ratio switching* for importance ratio control, and *discrepancy-aware advantage scaling* to preserve and amplify informative learning signals.

3.1 PROJECTED BEHAVIOR ESTIMATION

In importance sampling, the weight $r = \frac{\pi_\theta(o|q)}{\mu(o|q)}$ corrects for distribution shift, but the behavior distribution μ is often unknown. To obtain a tractable substitute, we project μ onto the target policy class Θ, yielding a parametric approximation evaluable within the policy family. We obtain the *projected behavior policy* (or *proxy policy*) $\pi_{\text{proj}} = \pi_{\hat{\phi}}$ by approximately minimizing the empirical cross-entropy objective:

$$\mathcal{L}_{\text{proj}}(\phi) = -\frac{1}{|\mathcal{D}_{\text{off}}|} \sum_{(q,o)\in\mathcal{D}_{\text{off}}} \log \pi_\phi(o|q), \qquad \phi \in \Theta. \tag{3}$$

Here $\hat{\phi}$ denotes the parameters obtained by supervised training on all trajectories in $\mathcal{D}_{\text{off}}$, including both positive and negative samples. This empirical projection provides a tractable approximation of behavior support within the target model family. This gives the proxy ratio $r_{\text{proj}} = \frac{\pi_\theta(o|q)}{\pi_{\text{proj}}(o|q)}$.

Because the external trajectories may come from stronger or heterogeneous policies, this projection is inevitably imperfect. We denote the approximation error by $\epsilon(o|q) = \mu(o|q)/\pi_{\text{proj}}(o|q)$ (Appendix C.1). When π_{proj} underestimates the support of an observed trajectory, the proxy ratio can be artificially inflated; when it overestimates the support, useful updates can be overly suppressed. Thus, π_{proj} gives a useful behavior-aware correction signal, but should not be trusted uniformly.

3.2 ADAPTIVE RATIO SWITCHING

To mitigate the risk of estimator variance described above, we propose an *adaptive ratio switching* mechanism that controls the surrogate ratio. We use the magnitude of the proxy ratio r_{proj} as a practical gating signal and define two candidate ratios:

Off-policy proxy ratio: $r_{\text{proj},t} = \frac{\pi_\theta(o_t|c_t)}{\pi_{\text{proj}}(o_t|c_t)}$. This ratio carries behavior-aware correction information, but can become unreliable when the projected denominator has large approximation error.

Local trust-region anchor: $r_{\text{safe},t} = \frac{\pi_\theta(o_t|c_t)}{\pi_{\theta_{\text{old}}}(o_t|c_t)}$. This ratio anchors the update to the previous policy iterate, the local reference used by clipped policy optimization. It does not estimate the external behavior distribution. We dynamically select the estimator based on whether the proxy ratio falls within a trusted range:

$$\hat{r}_t = \begin{cases} r_{\text{proj},t} & \text{if } |r_{\text{proj},t} - 1| \leq \epsilon_p \\ r_{\text{safe},t} & \text{otherwise} \end{cases}. \tag{4}$$

Interpretation. Adaptive switching combines behavior-aware correction with a local policy anchor. When the proxy ratio meets the gating criterion, APRO retains its off-policy correction; otherwise, it uses the previous-policy ratio to limit the influence of extreme proxy weights. Appendix C.2 analyzes the resulting ratio control under the trust-region assumption.

3.3 Discrepancy-Aware Advantage Scaling

After ratio switching controls the scale of the off-policy update, another issue remains: the update may still be too weak on the actions that matter most for capability transfer. In external reasoning trajectories, a rewarded token can have very low probability under the target policy precisely because it represents a behavior the weak model rarely discovers by itself. When the corresponding proxy ratio is small, its weighting can attenuate the positive learning signal on such tokens.

We measure this mismatch using an action-local discrepancy between the target policy's most preferred token and the demonstrated token: $\delta_t = \text{sg}\left[\frac{\max_{v\in\mathcal{V}} \pi_\theta(v|c_t)}{\pi_\theta(o_t|c_t)}\right]$.

A large δ_t indicates that the demonstrated action lies far from the target policy's current preference. For positive-advantage trajectories, this corresponds to a useful but under-explored action; for negative-advantage trajectories, it often means that the model already assigns low probability to the undesirable action.

To share the scaling signal across a trajectory, we aggregate token discrepancies as $\Delta_i = \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \delta_t$. We apply the discrepancy directly to the advantage in the nominal asymmetric form:

$$\tilde{A}_{i,t}^{\text{raw}} = \begin{cases} A_{i,t}\Delta_i, & A_{i,t} \geq 0, \\ A_{i,t}\Delta_i^{-1}, & A_{i,t} < 0. \end{cases} \tag{5}$$

The implemented scaled advantage $\tilde{A}_{i,t}$ uses the detached, clipped scaling specified in Appendix E. For $A_{i,t} > 0$, a large Δ_i marks a rewarded trajectory that the target policy currently fails to support, so the update is amplified to counteract signal attenuation. For $A_{i,t} < 0$, a large Δ_i means the negative trajectory is already far from the model's preferred action, so the penalty is down-weighted to avoid spending update budget on actions the model is unlikely to take. Therefore, discrepancy-aware scaling is tied to the same off-policy mismatch addressed by ratio switching: switching controls unreliable denominators, while scaling controls which reward-labeled mismatches should still produce strong learning signals.

3.4 The APRO Objective

The final objective combines the two parts of APRO's off-policy correction: $\hat{r}_t$ determines how much the update should trust the projected behavior denominator, while $\tilde{A}_t$ determines how strongly a reward-labeled mismatch should affect the policy. This yields a general off-policy objective compatible with standard policy gradient methods (e.g., PPO or GRPO):

$$\mathcal{J}(\theta) = \mathbb{E}_{(q,o)\sim\mathcal{D}_{\text{off}}}\left[\frac{1}{T}\sum_{t=1}^{T} \min\left(\hat{r}_t\tilde{A}_t, \text{clip}(\hat{r}_t, 1-\epsilon, 1+\epsilon)\tilde{A}_t\right)\right] - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}), \tag{6}$$

where the advantage $\tilde{A}_t$ can be defined at the token level as in PPO or at the sequence level as in GRPO. The resulting clipped surrogate combines proxy correction, adaptive switching, and discrepancy-aware weighting to learn from static external trajectories.

4 Experimental Setup

We evaluate APRO across two distinct domains: *mathematical reasoning*, which is characterized by objectively verifiable rewards, and *general instruction following*, which relies on preference-based reward models.

4.1 Datasets and Benchmarks

Mathematical Reasoning We construct a high-quality off-policy dataset by first using the subset of OpenR1-Math-220k Hugging Face (2025) selected by Yan et al. (2025) as prompts. We then sample reasoning trajectories for these prompts from three strong models: DeepSeek-R1 DeepSeek-AI (2025), QwQ-32B Qwen Team (2025), and DeepSeek-R1-Distill-Qwen-32B DeepSeek-AI (2025).

To ensure data quality and training stability, we apply the following data filters: (1) *Length*: Trajectories exceeding 8,192 tokens are excluded. (2) *Balance*: We retain only prompts containing both correct (positive) and incorrect (negative) responses to provide effective contrastive signals. The resulting dataset comprises 4,746 prompts, each associated with eight off-policy rollout trajectories.

For evaluation, we report *avg@32* for AIME 2024 MAA (2024), AIME 2025 MAA (2025), and AMC MAA (2023) to account for their limited data sizes. For larger-scale benchmarks, including Minerva Lewkowycz et al. (2022), OlympiadBench He et al. (2024), and MATH-500 Hendrycks et al. (2021), we report standard *pass@1* accuracy. Furthermore, to assess the robustness of the learned policy, we evaluate on out-of-distribution (OOD) tasks, including ARC-c Clark et al. (2018) and MMLU-Pro Wang et al. (2024b).

General Instruction Following To evaluate general alignment, we leverage the UltraFeedback dataset Cui et al. (2023) containing approximately 64k prompts. To construct the off-policy data, we follow the established protocols Cui et al. (2023); Meng et al. (2024); Chen et al. (2025) by sampling responses from a diverse ensemble of frontier models (i.e., Gemma-2-27B-It Gemma Team et al. (2024), Mistral-Large-Instruct-2407, Qwen-2.5-72B-Instruct Yang et al. (2024a), and Llama-3.1-70B-Instruct Dubey et al. (2024)). These responses are scored using ArmoRM-Llama3-8B-v0.1 Wang et al. (2024a) as the reward model. We assess performance on two standard benchmarks: AlpacaEval-2 Dubois et al. (2024) and Arena-Hard v2.0 Li et al. (2024). For AlpacaEval-2, we report both the length-controlled (LC) and raw win rates (WR). For Arena-Hard v2.0, we report win rates on both the hard-prompt and creative-writing task.

4.2 Baselines

We evaluate APRO against a diverse set of baselines, ranging from SFT to various RL algorithms (i.e., RWR Peters & Schaal (2007), AWR Peng et al. (2019), NLL-UL Welleck et al. (2019), ASPO Wang et al. (2025), DCPO Yang et al. (2025), BAPO Xi et al. (2025), and TOPR Roux et al. (2025)). All fully off-policy methods use the same external trajectories and rewards under the *black-box* constraint. Applicable ratio-based baselines use either $\pi_\theta/\pi_{\text{proj}}$ or the local anchor $\pi_\theta/\pi_{\text{old}}$. We also compare with value-based TBRM (Yuan et al., 2025) and joint PPO/GRPO+SFT objectives, and include PPO(On) and GRPO(On), trained on the same prompts via on-policy rollouts. Details are provided in Appendix D.

4.3 Implementation Details

Backbone Models We employ Qwen-2.5-Math-1.5B Yang et al. (2024b) and DeepSeek-R1-Distill-Qwen-1.5B as π_θ for mathematical reasoning, while Llama-3.2-3B-Instruct Meta AI (2024) is used for general instruction following. The projected behavior policy (π_{proj}) is initialized from the same base model as π_θ and trained on the complete off-policy dataset using SFT before the RL phase.

Training Settings We use `Math-Verify` binary correctness rewards for mathematical reasoning and ArmoRM scores for instruction following, with the `verl` framework (Sheng et al., 2024). The training and update batch sizes are 128 and 32. Fully off-policy training iterates over $\mathcal{D}_{\text{off}}$ without generating new rollouts. Sequence lengths, optimization settings, and stage-wise costs are given in Appendix E.

5 Results and Analysis

5.1 Main Results

Mathematical Reasoning Table 1 compares APRO with imitation and RL baselines on two mathematical reasoning backbones. On Qwen-2.5-Math-1.5B, APRO(PPO) reaches 35.5 ID average, compared with 31.0 for SFT, 32.4–32.8 for RWR/AWR/NLL-UL, and 28.3 for TBRM. On DeepSeek-R1-Distill-Qwen-1.5B, APRO(GRPO) reaches 45.3, compared with 41.9 for the strongest regression-style baseline and 44.4 for TBRM. Joint supervised assistance also improves the baselines: PPO+SFT

Table 1: Performance comparison on mathematical reasoning benchmarks. **Bold** indicates the best result, and underlined indicates the second best. Ratio-based and joint RL+SFT baselines report *X/Y*: proxy ratio ($\pi_\theta/\pi_{\text{proj}}$) / local-anchor ratio ($\pi_\theta/\pi_{\text{old}}$).

Method	In-Distribution Performance							Out-of-Distribution Performance		
	AIME 24	AIME 25	AMC	MATH-500	Minerva	Olympiad	Avg.	ARC-c	MMLU-Pro	Avg.
Qwen-2.5-Math-1.5B										
Base	7.4	3.0	23.3	34.0	9.6	18.7	16.0	2.6	4.2	3.4
Likelihood-based offline objectives										
SFT	11.1	11.4	36.2	68.6	22.8	36.1	31.0	36.2	19.1	27.7
RWR	12.8	13.9	37.9	69.4	27.6	34.2	32.6	32.2	20.6	26.4
AWR	11.6	14.1	37.9	72.0	27.2	34.2	32.8	25.5	18.8	22.2
NLL-UL	11.8	13.6	36.8	70.4	28.3	33.3	32.4	30.7	20.1	25.4
Value-based off-policy RL										
TBRM	9.8	7.9	33.7	65.0	26.4	26.7	28.3	28.2	22.3	25.3
Joint RL + SFT										
PPO+SFT	13.2/10.9	**14.2**/14.0	38.0/39.8	71.2/71.2	25.7/26.5	36.1/35.5	33.1/33.0	29.3/30.1	22.9/23.1	26.1/26.6
GRPO+SFT	11.5/10.0	12.7/9.3	36.6/39.0	68.2/72.0	25.7/26.5	32.4/35.4	31.2/32.0	32.3/28.8	19.1/24.6	25.7/26.7
Ratio-based off-policy RL										
TOPR	12.5/9.3	13.8/7.5	39.0/32.9	70.6/65.8	29.0/23.9	36.0/32.5	33.5/28.7	35.6/25.9	23.5/13.1	29.6/19.5
ASPO	5.6/7.8	8.4/4.0	27.6/32.2	56.0/55.2	17.6/17.3	25.0/25.7	23.4/23.7	39.8/1.1	16.0/6.5	27.9/3.8
DCPO	10.5/7.5	8.3/4.3	36.1/29.2	63.0/58.0	14.3/19.9	28.7/23.6	26.8/23.8	6.8/12.9	4.7/15.5	5.8/14.2
BAPO	13.3/7.6	9.1/4.5	39.9/32.6	70.4/55.8	27.2/14.7	34.5/26.3	32.4/23.6	35.4/2.9	19.5/9.7	27.5/6.3
PPO	10.5/5.6	11.7/2.7	**41.9**/27.3	73.8/47.8	30.1/21.3	36.1/24.4	34.0/21.5	34.7/7.0	23.6/15.3	29.2/11.2
GRPO	12.0/6.7	7.3/3.5	39.5/31.1	69.0/47.4	23.2/16.2	33.7/22.4	30.8/21.2	26.6/1.0	17.4/8.0	22.0/4.5
APRO(PPO)	**13.8**	13.6	41.0	**74.4**	**31.3**	**38.9**	**35.5**	40.2	24.7	32.5
APRO(GRPO)	12.3	10.3	39.5	71.2	27.9	37.7	33.2	40.5	22.3	31.4
On-policy references										
PPO(On)	12.8	6.4	41.6	71.4	26.5	33.0	32.0	47.4	**29.8**	**38.6**
GRPO(On)	13.1	6.1	39.3	66.8	27.2	31.5	30.7	**48.0**	23.7	35.9
DeepSeek-R1-Distill-Qwen-1.5B										
Base	19.0	16.7	51.2	72.8	23.5	40.0	37.2	5.7	18.4	12.1
Likelihood-based offline objectives										
SFT	20.5	17.4	55.1	81.4	30.9	45.2	41.8	18.8	19.5	19.2
RWR	23.0	20.8	54.7	79.4	27.9	43.5	41.6	11.0	14.9	13.0
AWR	24.3	19.4	54.6	77.8	31.3	43.8	41.9	10.8	14.1	12.5
NLL-UL	21.6	19.5	52.7	80.8	30.1	43.4	41.4	11.5	12.9	12.2
Value-based off-policy RL										
TBRM	23.4	20.4	58.8	83.2	32.0	48.7	44.4	25.4	21.6	23.5
Joint RL + SFT										
PPO+SFT	22.2/23.1	19.8/20.7	53.4/57.8	82.6/82.4	28.7/32.3	45.6/48.4	42.0/44.1	19.6/17.6	19.8/20.8	19.7/19.2
GRPO+SFT	22.9/21.6	20.4/19.6	53.5/56.4	79.6/81.6	31.3/34.6	42.9/46.0	41.8/43.3	24.0/19.8	19.6/26.2	21.8/23.0
Ratio-based off-policy RL										
TOPR	24.3/22.1	20.7/22.0	57.3/58.3	83.2/79.8	33.1/32.7	47.5/48.3	44.4/43.9	24.9/27.0	23.2/23.9	24.1/25.5
ASPO	20.7/22.3	19.8/19.5	51.4/60.2	75.8/79.0	28.3/29.0	41.3/46.4	39.6/42.7	42.9/22.6	21.4/20.7	32.2/21.7
DCPO	21.6/22.0	20.6/20.0	57.7/60.5	79.8/80.4	**35.7**/27.6	46.8/48.0	43.7/43.1	32.1/9.6	24.3/12.5	28.2/11.1
BAPO	21.4/22.1	18.1/18.2	49.5/60.1	77.0/83.2	27.9/35.3	41.5/48.7	39.2/44.6	27.6/13.9	18.5/23.3	23.1/18.6
PPO	19.4/22.1	20.3/18.3	49.5/**60.9**	77.0/80.6	27.9/30.1	39.1/49.3	38.9/43.6	24.2/8.4	18.0/12.3	21.1/10.4
GRPO	21.8/20.1	19.5/18.2	50.2/58.1	75.0/78.4	26.1/27.9	41.8/47.8	39.1/41.8	27.8/8.0	19.2/18.0	23.5/13.0
APRO(PPO)	**25.8**	**22.7**	58.1	82.6	29.8	49.8	44.8	28.2	21.3	24.8
APRO(GRPO)	23.1	22.2	59.9	**84.0**	30.9	**51.6**	**45.3**	36.6	**29.5**	33.1
On-policy references										
PPO(On)	22.7	19.8	54.9	78.6	25.4	45.6	41.2	**47.3**	27.6	**37.5**
GRPO(On)	20.3	18.3	54.8	80.2	26.8	45.8	41.0	35.1	26.9	31.0

reaches 33.1 on Qwen and 44.1 on DeepSeek, versus APRO(PPO)'s 35.5 and 44.8. GRPO+SFT reaches 32.0 and 43.3, versus APRO(GRPO)'s 33.2 and 45.3. APRO's gains thus extend beyond adding a supervised term to the RL objective.

OOD performance. For OOD evaluation, APRO(GRPO) achieves 31.4 on Qwen-2.5-Math-1.5B and 33.1 on DeepSeek-R1-Distill-Qwen-1.5B, outperforming most black-box off-policy and distillation baselines. APRO(PPO) further reaches 32.5 on Qwen. Pure on-policy PPO obtains higher OOD averages on both backbones, highlighting the complementary role of fully off-policy learning as a capability-transfer stage. Table 7 in Appendix H shows that continuing APRO(GRPO) checkpoints with on-policy GRPO raises the OOD averages to 39.5 and 35.9, respectively. These gains support APRO as a cold-start stage that prepares weak policies for further self-exploration.

Table 2: Results of APRO with other baselines on AlpacaEval-2 and Arena-Hard v2.0. Results are evaluated using GPT-4o-mini.

Method	AlpacaEval-2			Arena-Hard v2.0	
	LC(%)	WR(%)	Avg. Len.	Hard(%)	Creative(%)
Llama-3.2-3B-Instruct	29.7	33.9	2095	0.8	5.8
SFT	35.2	32.1	1899	0.8	5.7
PPO(proxy)	29.1	25.5	1817	1.1	3.9
PPO(local)	33.8	**38.8**	2072	1.2	6.1
GRPO(proxy)	29.6	16.3	1060	0.8	2.1
GRPO(local)	35.8	27.0	1350	0.9	3.6
APRO(PPO)	36.7	33.2	1873	**1.5**	5.7
APRO(GRPO)	**39.7**	35.4	1876	1.3	**6.6**

Instruction Following We then evaluate APRO on general instruction following benchmarks (Table 2), where it demonstrates remarkable efficacy and stability, consistent with its performance in mathematical reasoning. ***First***, APRO breaks the imitation ceiling by achieving a leading 39.7% LC win rate on AlpacaEval-2, exceeding SFT (35.2%) and underscoring the value of off-policy feedback. ***Second***, APRO effectively mitigates off-policy volatility. Unlike proxy-based baselines that suffer from performance degradation (e.g., GRPO(proxy) drops sharply in raw win rate from 33.9% to 16.3%), APRO maintains stable training and achieves higher length-controlled win rates than the local-anchor baselines. ***Third***, APRO excels in handling complex queries. On the challenging Arena-Hard v2.0 benchmark, APRO(PPO) achieves a 1.5% win rate on hard prompts, while APRO(GRPO) achieves a 6.6% win rate on creative writing prompts, outperforming most baselines in tackling difficult instructions.

5.2 Ablation Study

We perform ablation studies on the mathematical reasoning task to assess the contribution of each core component, using DeepSeek-R1-Distill-Qwen-1.5B as the backbone model.

Adaptive Ratio Switching Table 3 shows that the pure proxy baseline reaches 39.1 ID, compared with 41.8 for the local-anchor baseline. Adding adaptive switching raises the proxy baseline to 41.1. Figure 2(a) shows the accompanying reduction in clipping activity, consistent with restricting extreme proxy ratios through the local fallback.

Discrepancy-Aware Scaling Table 3 shows that combining scaling with switching achieves the best ID/OOD results, 45.3/33.1. Switching alone reaches 41.1/22.9, while scaling with the local denominator reaches 42.9/25.6; neither reproduces the full OOD result. Figure 2(a) also shows a lower *Switch Rate* for APRO (green) than for the variant without scaling (purple), alongside the controlled entropy dynamics in Figure 2(b). These observations support complementary effects in this configuration.

Table 3: Ablation studies on mathematical reasoning. "Switching" refers to *adaptive ratio switching* (ARS), and "Scaling" refers to *discrepancy-aware advantage scaling*. The "Ratio Base" column indicates the denominator used in importance sampling. *ID* means in-distribution and *OOD* means out-of-distribution performance.

Switching	Scaling	Ratio Base	ID (Avg.)	OOD (Avg.)
✗	✗	π_{old}	41.8	13.0
✗	✗	π_{proj}	39.1	23.5
✓	✗	ARS	41.1	22.9
✗	✓	π_{old}	42.9	25.6
✗	✓	π_{proj}	36.8	25.8
✓	✓	ARS	45.3	33.1

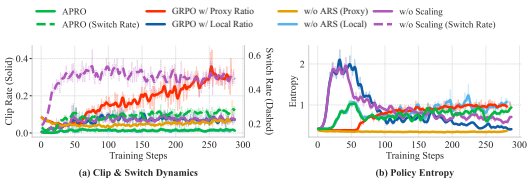

Figure 2: Training dynamics comparing APRO and its variants. **(a)** Clip and switch rates. **(b)** Entropy across training steps.

Additional analysis. We further study the projected behavior policy and oracle behavior probabilities in Appendix F and Appendix G. The results show that replacing π_{proj} with random or constant denominators substantially hurts performance, confirming the importance of behavior-aware support information. The complete APRO update also improves performance with oracle behavior probabilities. Appendix I.2 separately studies projector training epochs and the use of positive/negative trajectory subsets.

Switching-threshold sensitivity. Table 4 shows that ID performance changes little for $\epsilon_p \in [0.01, 0.2]$, whereas OOD performance is more sensitive. The default 0.1 is competitive with 0.05; both overly restrictive and permissive gates reduce OOD scores. The threshold is fixed during each run, while the switching decision is token-adaptive.

5.3 Analysis of Training Stability

A key motivation behind APRO is to address the trade-off between variance explosion and signal attenuation. Figure 3 illustrates the training dynamics; Appendix I.1 reports windowed gradient-norm statistics for APRO.

Mitigating Signal Attenuation In the proxy regime, GRPO and other proxy-ratio baselines exhibit persistently small or poorly controlled gradient norms, highlighting a critical issue: informative

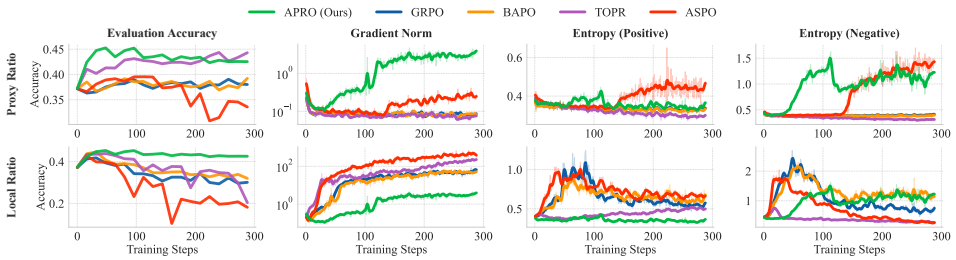

Figure 3: Training dynamics of APRO against different baselines across evaluation accuracy, gradient norms, and policy entropy on training data. **Top**: pure off-policy proxy estimator ($r = \pi_\theta/\pi_{\text{proj}}$). **Bottom**: local-anchor estimator ($r = \pi_\theta/\pi_{\text{old}}$).

off-policy tokens often reside in the low-probability tail of the target policy, leading to near-zero ratios and suppressing the learning signal. APRO resolves this challenge through discrepancy-aware scaling, which amplifies these weak yet essential signals, restoring robust gradient magnitudes and enabling sustained learning where conventional baselines fail.

Mitigating Estimation Bias in Conservative Updates In the conservative setting, a clear pattern emerges: methods relying on the local-anchor ratio ($\pi_\theta/\pi_{\text{old}}$) show a brief initial improvement, followed by degradation or unstable convergence. This pattern is consistent with bias under behavior mismatch. Specifically, the estimator treats off-policy data as if it was generated by the target policy, neglecting the distribution shift between the behavior policy and the target policy. Although the importance ratios appear bounded near 1.0, giving a false sense of stability, the resulting gradient updates are misaligned, causing the model to naively mimic external trajectories without proper importance weighting. Thus, proximity to the previous policy does not itself correct behavior mismatch or ensure performance retention.

Table 4: Switching-threshold sensitivity for APRO(GRPO) on DeepSeek-R1-Distill-Qwen-1.5B. The surrogate clipping width remains $\epsilon = 0.1$.

ϵ_p	ID Avg.	OOD Avg.
0.01	45.1	22.1
0.05	45.4	34.8
0.10 (default)	45.3	33.1
0.20	45.2	24.6
0.50	41.8	15.9

Dynamics of Policy Entropy The entropy dynamics show that APRO maintains a clearer separation between positive and negative trajectories. On positive samples, APRO keeps the entropy low and stable, suggesting that the model learns to assign confident probability mass to rewarded reasoning paths. On negative samples, APRO maintains a higher but controlled entropy, preventing the policy from becoming overconfident on undesirable trajectories. This behavior contrasts with baselines whose entropy either grows unstably, such as ASPO, or remains poorly aligned with the accuracy trend under different denominator choices. Combined with the gradient norm curves, the results indicate that APRO stabilizes off-policy learning by reinforcing positive trajectories while avoiding overly aggressive suppression of negative ones.

6 Conclusion

We studied fully off-policy RL for LLM post-training to address weak policies' limited self-exploration. APRO combines projected behavior estimation, adaptive ratio switching, and discrepancy-aware advantage scaling to regulate off-policy ratios and strengthen informative updates. Experiments on mathematical reasoning and instruction following show stable training and improved performance over the evaluated imitation and ratio-based baselines. Further results demonstrate the value of external trajectories in mixed-policy training and of APRO checkpoints as initializations for subsequent on-policy RL. These findings support fully off-policy learning as a useful capability-transfer stage for weak target policies.

AI USE STATEMENT

We used large language models as a writing assistant to polish phrasing and improve the readability of the manuscript, as well as for minor code refactoring (e.g., formatting and commenting). All technical content, claims, experiments, and conclusions were produced and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Ruijun Chen, Jiajian Guo, Hongzhan Chen, Fanqi Wan, Qifan Wang, and Xiaojun Quan. ReAlign: Structured revision for small language model alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12005–12020, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li, and Ge Li. Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. *arXiv preprint arXiv:2508.00222*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Weixin Xu, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

MAA. American mathematics competitions, 2023. URL https://www.maa.org/math-competitions. Online competition series.

MAA. American invitational mathematics examination (aime), 2024. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

MAA. American invitational mathematics examination (aime), 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, September 2024. 15 minute read.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever,

Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on Machine learning*, pp. 745–750, 2007.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Nicolas Le Roux, Marc G Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy reinforce: Stable and efficient reinforcement learning for llms. *arXiv preprint arXiv:2503.14286*, 2025.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1889–1897. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/schulman15.html.

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

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

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

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024a.

Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv: 2504.14945*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

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

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yurun Yuan, Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Trajectory bellman residual minimization: A simple value-based method for LLM reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 116994–117032. Curran Associates, Inc., 2025. doi: 10.52202/085713-3900. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a9a3c5df70368a2f8c3fd2ed12620203-Paper-Conference.pdf.

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

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*, 2025.

Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, Pengfei Zheng, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Ruiming Tang, Shiyao Wang, Shujie Yang, Tao Wu, Wuchao Li, Xinchen Luo, Xingmei Wang, Yi Su, Yunfan Wu, Zexuan Cheng, Zhanyu Liu, Zixing Zhang, Bin Zhang, Boxuan Wang, Chaoyi Ma, Chengru Song, Chenhui Wang, Chenglong Chu, Di Wang, Dongxue Meng, Dunju Zang, Fan Yang, Fangyu Zhang, Feng Jiang, Fuxing Zhang, Gang Wang, Guowang Zhang, Han Li, Honghui Bao, Hongyang Cao, Jiaming Huang, Jiapeng Chen, Jiaqiang Liu, Jinghui Jia, Kun Gai, Lantao Hu, Liang Zeng, Qiang Wang, Qidong Zhou, Rongzhou Zhang, Shengzhe Wang, Shihui He, Shuang Yang, Siyang Mao, Sui Huang, Tiantian He, Tingting Gao, Wei Yuan, Xiao Liang, Xiaoxiao Xu, Xugang Liu, Yan Wang, Yang Zhou, Yi Wang, Yiwu Liu, Yue Song, Yufei Zhang, Yunfeng Zhao, Zhixin Ling, and Ziming Li. Onerec-v2 technical report. *arXiv preprint arXiv:2508.20900*, 2025.

A Limitations

APRO is designed for black-box off-policy training with static external trajectories, and its effectiveness depends on whether these trajectories provide useful support for the target task. When the off-policy data is extremely mismatched with the target policy or contains little reward diversity, the projected behavior policy may provide a weaker correction signal, and the switching mechanism may fall back to the local trust-region anchor more frequently. Our analysis of projected behavior policies and oracle probabilities suggests that better behavior support improves stability, but accurate denominators alone are not sufficient to remove signal attenuation.

Another limitation is that APRO introduces several interacting components. Although our ablations show that projected behavior estimation, adaptive ratio switching, and discrepancy-aware advantage scaling address complementary failure modes, the current design is not necessarily the only way to combine behavior correction and signal preservation. Developing simpler objectives with similar stability remains an interesting direction for future work.

B Related Work

RL for reasoning and the limitation of self-exploration. Reinforcement learning has become an important post-training technique for improving the reasoning ability of LLMs. Recent systems such as OpenAI o1, DeepSeek-R1, and Kimi k1.5 show that RL with verifiable rewards can substantially improve long-chain reasoning and mathematical problem solving (OpenAI et al., 2024; DeepSeek-AI, 2025; Kimi Team et al., 2025). Open-source RLVR efforts further demonstrate that scalable on-policy training can be effective when the model is already capable of sampling informative trajectories (Hu et al., 2025; Yu et al., 2025). However, purely on-policy self-exploration is constrained by the current policy's own support. When the initial model is weak, correct trajectories can be rare, and RL may mostly reinforce behaviors that already exist in the model distribution. Recent analyses also suggest that RL post-training can amplify pre-existing behaviors rather than reliably create new capabilities from scratch (Zhao et al., 2025; Yue et al., 2025). Our work focuses on this weak-policy regime, where external trajectories are useful because the target policy may not be able to discover high-quality solutions by itself.

Learning from external trajectories: imitation and regression-style objectives. A direct way to use external trajectories is supervised fine-tuning or knowledge distillation. Such objectives are simple and strong, and they provide an important cold-start baseline in our experiments. Several related methods further incorporate reward or preference information into likelihood-style training. Reward-weighted regression (RWR) weights the likelihood objective by reward (Peters & Schaal, 2007), while advantage-weighted regression (AWR) interprets off-policy RL as weighted supervised regression (Peng et al., 2019). In language generation, unlikelihood training penalizes undesirable tokens and provides a natural way to use negative samples (Welleck et al., 2019). Self-training and filtering-based pipelines, such as ReST and ReST-EM-style approaches, also use generated trajectories and rewards to iteratively improve language models (Gulcehre et al., 2023; Singh et al., 2023). These methods are closely related to our setting because they can all learn from static external data. However, they remain primarily imitation or regression-style objectives: they increase or decrease trajectory likelihood according to reward signals, but do not explicitly correct the mismatch

between the target policy and the behavior policy through an off-policy policy-gradient ratio. APRO is designed to test whether reward-driven off-policy policy optimization can provide gains beyond this imitation ceiling.

On-policy and clipping-based policy optimization. PPO and TRPO stabilize policy-gradient learning by constraining the update around a recent behavior policy (Schulman et al., 2017; 2015). Recent LLM post-training methods build on clipped policy optimization with mechanisms designed to alleviate off-policy instability or ratio imbalance. ASPO introduces asymmetric importance sampling to handle imbalance between positive and negative samples (Wang et al., 2025); DCPO dynamically adjusts clipping behavior under changing optimization conditions (Yang et al., 2025); BAPO studies instability caused by off-policy negative samples and proposes balanced adaptive clipping (Xi et al., 2025); and GBPO further explores group-based policy optimization for LLM training (Zhou et al., 2025). These methods are therefore natural policy-gradient baselines for APRO, since they also address instability arising from importance ratios, clipping, or off-policy samples. However, they are not primarily evaluated in the fully static black-box off-policy regime considered in this paper, where all trajectories are generated by external policies and the true behavior probabilities are unavailable. We therefore adapt them to the same constraint as APRO, using either a projected proxy denominator or a local trust-region denominator, to test whether their existing stabilization mechanisms are sufficient in the fully off-policy setting.

Off-policy and hybrid-policy RL for LLMs. A growing line of work studies how to incorporate trajectories that are not sampled from the current policy. LUFFY uses off-policy guidance to help reasoning models learn from stronger trajectories beyond their own exploration distribution (Yan et al., 2025). RL-PLUS studies hybrid-policy optimization and argues that external rollouts can help counter capability-boundary collapse in RL training (Dong et al., 2025). TOPR proposes a tapered off-policy REINFORCE objective for stable and efficient LLM training from off-policy data (Roux et al., 2025). TBRM (Yuan et al., 2025) offers a value-based route through trajectory Bellman residual minimization, supporting offline learning without behavior importance ratios or a separate critic. These works support the general motivation that off-policy data can be useful for LLM post-training. APRO differs in its focus on the black-box fully off-policy regime: the external trajectories are static, the behavior logits are unavailable, and the target model may be much weaker than the behavior policies. Under this constraint, the key difficulty is not only how to use off-policy data, but how to construct a usable denominator, regulate unreliable ratios, and avoid suppressing informative low-probability actions.

C Theoretical Analysis

We analyze three properties of the APRO surrogate: error from the projected denominator, conditional control of the selected ratio, and discrepancy-based reweighting of local policy updates.

C.1 Bias-Variance Analysis of the Proxy Estimator

Problem Setup. Let x denote a complete response to a prompt drawn from a fixed prompt distribution, $\mu(x)$ the distribution of retained off-policy responses, and $\pi_{\text{proj}}(x)$ the projected proxy policy. Assume π_θ is supported by μ, μ is supported by π_{proj}, and the following expectations exist. For a response-level weight $A(x)$ held fixed during differentiation, the trajectory-level importance-sampling (IS) surrogate gradient is

$$g_\mu = \mathbb{E}_{x\sim\mu}\left[\frac{\pi_\theta(x)}{\mu(x)}\nabla_\theta \log \pi_\theta(x)A(x)\right]. \tag{7}$$

Substituting the available proxy denominator gives

$$g_{\text{proxy}} = \mathbb{E}_{x\sim\mu}\left[\frac{\pi_\theta(x)}{\pi_{\text{proj}}(x)}\nabla_\theta \log \pi_\theta(x)A(x)\right]. \tag{8}$$

Theorem C.1 (Effect of the Proxy Denominator)**.** *The difference between these surrogate gradients is*

$$g_{proxy} - g_\mu = \mathbb{E}_{x\sim\pi_\theta}\left[\left(\frac{\mu(x)}{\pi_{proj}(x)} - 1\right)\nabla_\theta \log \pi_\theta(x)A(x)\right]. \tag{9}$$

Proof. Changing the measure in the proxy expectation gives

$$g_{\text{proxy}} = \int \mu(x) \frac{\pi_\theta(x)}{\pi_{\text{proj}}(x)} \nabla_\theta \log \pi_\theta(x) A(x) dx \tag{10}$$

$$= \int \pi_\theta(x) \left(\frac{\mu(x)}{\pi_{\text{proj}}(x)} \right) \nabla_\theta \log \pi_\theta(x) A(x) dx. \tag{11}$$

The ideal surrogate is $g_\mu = \int \pi_\theta(x) \nabla_\theta \log \pi_\theta(x) A(x)\, dx$. Subtracting yields the identity. The approximation error ratio $\rho(x) = \mu(x)/\pi_{\text{proj}}(x)$ therefore isolates the effect of denominator substitution for complete trajectories. Below, we analyze the token-conditional clipped surrogate used by APRO.

Proxy-Ratio Second Moment. Let $G_{\text{proxy}}(x)$ be the single-response quantity inside the proxy expectation, with vector variance defined as $\mathrm{Var}(G) = \mathbb{E}\|G - \mathbb{E}G\|^2$. Then

$$\mathrm{Var}(G_{\text{proxy}}) \le \mathbb{E}_{x\sim\mu} \left[\left(\frac{\pi_\theta(x)}{\pi_{\text{proj}}(x)} \right)^2 \|A(x) \nabla_\theta \log \pi_\theta(x)\|^2 \right]. \tag{12}$$

The proxy weight factors as $(\pi_\theta/\mu)(\mu/\pi_{\text{proj}})$. Proxy under-support can enlarge this weighted second moment when target probability and the advantage-weighted score remain appreciable. This motivates controlling extreme proxy ratios during policy optimization.

C.2 Variance Bounds via Adaptive Switching

We quantify how adaptive switching controls the selected token ratio under a local-anchor condition. Recall the switched ratio:

$$\hat{r}_t = \begin{cases} r_{\text{proj},t}, & |r_{\text{proj},t} - 1| \le \epsilon_p, \\ r_{\text{safe},t}, & |r_{\text{proj},t} - 1| > \epsilon_p, \end{cases} \qquad r_{\text{safe},t} = \frac{\pi_\theta(o_t|c_t)}{\pi_{\text{old}}(o_t|c_t)}. \tag{13}$$

Assumption C.2 (Trust Region Stability)**.** *For the sampled tokens and parameter iterates under consideration, assume $r_{safe,t} \in [1-\epsilon, 1+\epsilon]$ for some $0 < \epsilon < 1$.*

Theorem C.3 (Conditional Bound for the Switched Ratio)**.** *Under the Trust Region Stability assumption, the selected ratio satisfies $\hat{r}_t \in [1-C, 1+C]$, where $C = \max(\epsilon, \epsilon_p)$. At differentiable points away from switching and clipping boundaries, with fixed denominators and detached advantage multipliers, if $\mathbb{E}[\tilde{A}_t^2 \|\nabla_\theta \log \pi_\theta(o_t|c_t)\|^2] < \infty$, the per-token clipped policy-surrogate gradient has a finite second moment.*

Proof. Consider the two cases for any sampled token:

1. **Case 1: Inside the proxy-ratio interval ($|r_{\text{proj},t} - 1| \le \epsilon_p$).** Here $\hat{r}_t = r_{\text{proj},t} \in [1-\epsilon_p, 1+\epsilon_p]$ by definition.
2. **Case 2: Outside the proxy-ratio interval ($|r_{\text{proj},t} - 1| > \epsilon_p$).** Here $\hat{r}_t = r_{\text{safe},t} \in [1-\epsilon, 1+\epsilon]$ by assumption.

Combining the cases gives $|\hat{r}_t| \le 1 + C$. The active policy-surrogate gradient contribution equals $\tilde{A}_t \hat{r}_t \nabla_\theta \log \pi_\theta(o_t|c_t)$, while a clipped-flat contribution is zero. Therefore, for the per-token contribution G_t,

$$\mathbb{E}\|G_t\|^2 \le (1+C)^2 \mathbb{E}\Big[\tilde{A}_t^2 \|\nabla_\theta \log \pi_\theta(o_t|c_t)\|^2 \Big] < \infty. \tag{14}$$

The gate controls ratio magnitude: it retains the projected ratio near unity and uses the previous-policy anchor elsewhere. This balances behavior-aware correction with local update control.

C.3 Analysis of Signal Attenuation and Rectification

We analyze local gradient contributions to the clipped surrogate. Denominators, advantages, and discrepancy multipliers are held fixed during differentiation, and the analysis excludes switching and clipping boundaries. This separates the effect of the multiplier from changes in the sampling distribution or the selected denominator.

C.3.1 Signal Attenuation in Standard Clipped Objectives

Let $p_t = \pi_\theta(o_t|c_t)$, $b_t > 0$ be the selected fixed denominator, and $r_t = p_t/b_t$. Write $s_t = \nabla_\theta \log \pi_\theta(o_t|c_t)$. For

$$H(r_t, A_t) = \min\{r_t A_t, \mathrm{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t\}, \tag{15}$$

the local derivative is

$$\nabla_\theta H = \begin{cases} A_t r_t s_t, & A_t > 0,\ r_t < 1+\epsilon, \\ A_t r_t s_t, & A_t < 0,\ r_t > 1-\epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

In particular, for a positive advantage and a small ratio, the minimum selects the *unclipped* term. If b_t remains bounded away from zero and $A_t s_t$ is bounded, then $p_t \to 0$ implies $A_t(p_t/b_t)s_t \to 0$. The gradient contribution therefore decreases with the ratio. For a negative advantage with $r_t < 1-\epsilon$, the clipped objective already has zero local gradient.

This small-ratio effect characterizes fixed-proxy updates when the selected denominator remains appreciable. APRO combines ratio switching with discrepancy-aware scaling to regulate the contribution of such actions.

C.3.2 Mechanism of Discrepancy-Aware Rectification

For an observed token o_t, let $p_t^\star = \max_v \pi_\theta(v|c_t)$. Its detached discrepancy from the model's preferred token and the corresponding trajectory average are

$$\delta_t = \mathrm{sg}[p_t^\star/p_t], \qquad \Delta_i = \frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\delta_t. \tag{17}$$

The token discrepancy marks where the demonstrated action differs from the target policy's current preference. Averaging it across a trajectory shares this signal across its tokens. For a positive advantage, a large Δ_i increases the corresponding rewarded update; for a negative advantage, its inverse reduces the corresponding penalty.

For $0 < w_{\min} \le 1 \le w_{\max} < \infty$, define the detached multiplier and scaled advantage by

$$w_{i,t} = \mathrm{clip}\Big(\Delta_i^{\mathrm{sgn}(A_{i,t})}, w_{\min}, w_{\max}\Big), \qquad \tilde{A}_{i,t} = w_{i,t}A_{i,t}. \tag{18}$$

For GRPO, the sequence-level advantage shares one multiplier w_i across its tokens. For PPO, the sign follows each token-level advantage while Δ_i remains a trajectory average. Since $\Delta_i \ge 1$, positive contributions use $w_{i,t} \in [1, w_{\max}]$, and negative contributions use $w_{i,t} \in [w_{\min}, 1]$.

Theorem C.4 (Bounded Local Reweighting)**.** *Let g be an unscaled token-level clipped policy-surrogate gradient contribution and $\tilde{g}$ the contribution after applying its detached multiplier $w \in [w_{\min}, w_{\max}]$. Then*

$$\tilde{g} = wg, \qquad w_{\min}\|g\| \le \|\tilde{g}\| \le w_{\max}\|g\|. \tag{19}$$

If $\mathbb{E}\|g\| < \infty$, then

$$\|\mathbb{E}[\tilde{g}] - \mathbb{E}[g]\| \le \mathbb{E}[|w-1|\|g\|] \le \max\{1-w_{\min}, w_{\max}-1\}\,\mathbb{E}\|g\|. \tag{20}$$

If $\mathbb{E}\|g\|^2 < \infty$, then $\mathbb{E}\|\tilde{g}\|^2 \le w_{\max}^2\mathbb{E}\|g\|^2$.

Proof. For $w > 0$, $H(r, wA) = wH(r, A)$. Differentiating with w detached gives $\tilde{g} = wg$; the inequalities follow from the bounds on w and the triangle inequality.

With GRPO's shared multiplier, the same identity applies to the sum of token contributions in a trajectory. For PPO, where token-level advantage signs may differ, the bounds apply to individual token contributions.

D Baselines

We compare APRO with three families of baselines. The first family consists of *regression-style offline objectives*, which test whether weighted imitation over external trajectories is sufficient. This

group includes SFT, reward-weighted regression (RWR Peters & Schaal (2007)), advantage-weighted regression (AWR Peng et al. (2019)), and NLL-UL Welleck et al. (2019), which combines negative log-likelihood on preferred trajectories with unlikelihood training on dispreferred trajectories.

The second family consists of off-policy policy-gradient baselines. We include TOPR (Roux et al., 2025), PPO/GRPO-style objectives, and recent clipping-based variants such as ASPO Wang et al. (2025), DCPO Yang et al. (2025), and BAPO Xi et al. (2025). Since the true behavior probabilities of the external sources are unavailable, we evaluate applicable policy-gradient baselines under the same black-box constraint as APRO. Specifically, we instantiate two denominator choices: a projected proxy denominator π_{proj} and a local trust-region anchor $\pi_{\theta_{\text{old}}}$. These adapted settings correspond to the X/Y results in Table 1. We also evaluate online pipelines PPO(On) and GRPO(On), which generate rollouts from the current policy.

The third family is represented by TBRM (Yuan et al., 2025), which optimizes a trajectory-level Bellman residual without behavior importance ratios or a separate critic. We evaluate its static, mixed-policy, and continued-GRPO settings alongside the corresponding APRO settings.

We additionally evaluate joint PPO/GRPO+SFT objectives on positive external trajectories, with $L_{\text{hybrid}} = L_{\text{PPO/GRPO}} + \lambda L_{\text{SFT}}(\mathcal{D}^{+}_{\text{off}})$. We set $\lambda = 1$ in all joint PPO/GRPO+SFT experiments. These baselines use either the projected or local ratio in Table 1. The supervised and RL losses are optimized jointly; the SFT+RL setting in Table 7 instead uses an SFT checkpoint to initialize subsequent on-policy GRPO.

E Implementation Details

E.1 Projected Behavior Policy Training

The projected behavior policy π_{proj} serves as a proxy for the unknown behavior distribution μ. It is obtained by supervised fine-tuning from the same initialization as the target policy π_θ, using the static off-policy dataset $\mathcal{D}_{\text{off}}$. Unlike standard SFT baselines, which are trained only on correct solutions, π_{proj} is trained on all rollout trajectories in $\mathcal{D}_{\text{off}}$, including both correct and incorrect samples. This allows π_{proj} to better preserve the support structure of the full behavior distribution, rather than only modeling high-reward trajectories.

Mathematical Reasoning. For Qwen-2.5-Math-1.5B, we train π_{proj} for 5 epochs with a learning rate of 5×10^{-5} and a global batch size of 64. For DeepSeek-R1-Distill-Qwen-1.5B, we reduce the learning rate to 5×10^{-6} to avoid catastrophic forgetting of distilled reasoning behaviors, while keeping all other hyperparameters identical.

Instruction Following. For Llama-3.2-3B-Instruct, the projected policy is trained for 1 epoch using a learning rate of 5×10^{-5} and a global batch size of 64.

E.2 Reinforcement Learning and Baselines

We describe the training configurations for APRO and all baselines. Unless otherwise specified, all APRO variants and RL baselines in the main experiments are initialized from the corresponding base model, rather than from the SFT baseline. The projected behavior policy π_{proj} is only used to compute the proxy denominator for off-policy correction and is not used as the initialization of the optimized target policy. For the continued-RL experiments in Appendix H, methods marked with "+RL" are instead initialized from the corresponding first-stage checkpoint and then further optimized with on-policy GRPO.

In the main APRO experiments, the proxy threshold is $\epsilon_p = \epsilon = 0.1$; Table 4 varies only ϵ_p. The scaled advantage uses a detached, clipped discrepancy factor: $\tilde{A}_{i,t} = A_{i,t}\,\text{clip}(\Delta_i^{\text{sgn}(A_{i,t})}, w_{\min}, w_{\max})$, with $[w_{\min}, w_{\max}] = [0.1, 3.0]$. For PPO, the sign is taken from the corresponding token-level advantage while the discrepancy is aggregated over its trajectory.

Mathematical Reasoning. We set the maximum prompt length to 1,024 and the maximum response length to 8,192.

- **Qwen-2.5-Math-1.5B.** Following (Yan et al., 2025), we modify the model configuration by increasing the RoPE base from 10,000 to 40,000 and extending the sliding window size to 16,384 to support long-context reasoning. We train on $\mathcal{D}_{\text{off}}$ for 8 epochs with a learning rate of 1×10^{-5}, a KL coefficient of 0.01, and a PPO clipping range $\epsilon = 0.1$. All remaining hyperparameters follow the default settings of the corresponding baseline methods.
- **DeepSeek-R1-Distill-Qwen-1.5B.** We use the same training configuration as Qwen-2.5-Math-1.5B, except that the KL coefficient is increased to 0.1 and the clipping range remains $\epsilon = 0.1$.
- **SFT Baseline.** The SFT baseline is initialized from the base model and trained only on positive samples from $\mathcal{D}_{\text{off}}$ for 8 epochs, using a learning rate of 1×10^{-5}.

Instruction Following. We use a maximum sequence length of 2048 tokens.

- **APRO.** We train for 1 epoch with a learning rate of 5×10^{-6} and a KL coefficient of 0.1. The training batch size is set to 128 and the update batch size to 32.
- **RL Baselines.** All RL baselines are initialized from the base model. For these baselines, we use a smaller learning rate of 5×10^{-7} to ensure stable optimization, while keeping the remaining hyperparameters consistent with APRO.
- **SFT Baseline.** The SFT baseline is initialized from the base model and trained for 2 epochs with a learning rate of 1×10^{-6} and a global batch size of 128.

E.3 Computational Cost

Table 5 reports stage-wise training times on DeepSeek-R1-Distill-Qwen-1.5B using the same 64-GPU configuration. Projector training is a one-time cost for a fixed model and external dataset. The reported APRO totals include projector training and the RL stage; they are training-stage measurements, not an end-to-end accounting of external data collection. Relative to GRPO-based updates, APRO also evaluates the projected policy during its RL stage.

Table 5: Wall-clock training time under the same 64-GPU configuration.

Method or stage	Time (hours)
Projected behavior estimation	0.4
APRO (GRPO / PPO), RL stage	5.5 / 7.8
APRO (GRPO / PPO), total	5.9 / 8.2
RWR / AWR / NLL-UL / TOPR	5.0–5.2
ASPO / DCPO / BAPO / GRPO (off-policy)	5.2–5.5
PPO (off-policy)	7.3–7.8
GRPO (on-policy)	10.0
PPO (on-policy)	12.2

F Analysis of Projected Behavior Policy

We next examine the role of the projected behavior policy π_{proj}. A common concern is whether this component is truly necessary, since π_{proj} is itself obtained by supervised training on the off-policy dataset and is only an approximation to the unknown behavior distribution μ. Our goal is not to claim that π_{proj} exactly recovers μ. Instead, it provides a structured projection of the black-box behavior distribution into the target model family, so that the denominator in the importance ratio preserves meaningful information about the relative support of the observed trajectories. To test whether such a structured denominator is needed, we replace π_{proj} with two heuristic alternatives while keeping the rest of APRO, including discrepancy-aware scaling, unchanged: (1) Random Behavior, where $\pi_{\text{off}}(o_t|c_t)$ is sampled from $U(0,1)$ independently for each token; and (2) Constant Behavior, where $\pi_{\text{off}}(o_t|c_t) = 0.5$ for all tokens, making the ratio depend only on the target policy probability.

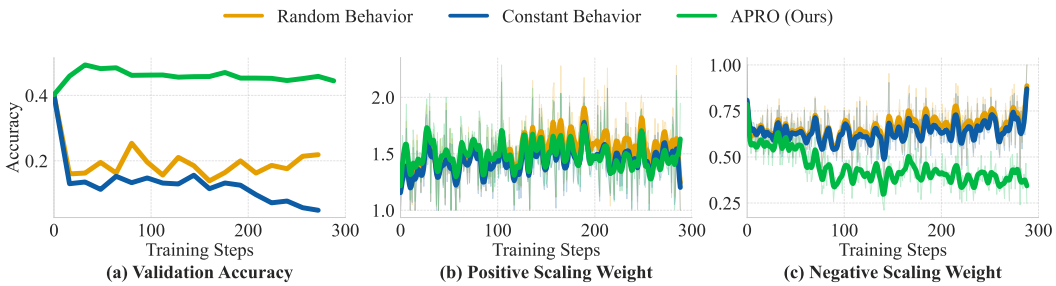

Figure 4: Analysis of the projected behavior policy. **(a)** Validation scores comparing APRO against heuristic baselines. **(b & c)** Positive and negative scaling weights during training.

Figure 4(a) shows that both heuristic denominators lead to severe degradation. The constant denominator quickly collapses after the first few updates, while the random denominator fails to produce stable improvement and remains far below APRO. This indicates that the benefit of APRO does not come merely from applying a clipped objective or rescaling advantages. The learned denominator provides behavior-support information absent from these heuristic alternatives and yields stronger validation performance under the same APRO optimization procedure.

The scaling dynamics in Figure 4(b,c) further characterize the effect of the denominator. Positive scaling weights fluctuate under all three choices, while APRO's negative scaling weight gradually decreases with the learned proxy. This trend is consistent with larger trajectory-level discrepancy and weaker weighting of negative updates. In contrast, the random and constant denominators retain larger, less stable negative weights. Together with the validation curves, these comparisons show the benefit of learned behavior-support information for coordinating ratio correction and discrepancy-aware scaling.

G Performance with Oracle Probabilities

While APRO addresses scenarios where behavior probabilities are unknown, a natural question emerges: *Does the proposed optimization framework remain advantageous when the true behavior probabilities are available?*

To investigate this, we utilize the DeepSeek-R1-Distill-32B model as an open-source behavior policy, allowing us to explicitly compute the true behavior probability $\mu(o|q)$. We compare two settings: (1) Standard IS (Oracle): using the true importance ratio $r = \pi_\theta/\mu$ without switching or scaling; (2) APRO (Oracle): using the same true ratio $r = \pi_\theta/\mu$, augmented with our switching and scaling mechanisms.

Table 6: Comparison with oracle behavior probabilities.

Method Setting	ID (Avg.)	OOD (Avg.)
Standard IS (Oracle)	38.6	9.9
APRO (Oracle)	43.6	32.3

As detailed in Table 6, Standard IS reaches 38.6% ID and 9.9% OOD with oracle probabilities, while APRO reaches 43.6% and 32.3%. The improvement highlights the value of regulating update weights even when the behavior denominator is known. For positive-advantage tokens with $\pi_\theta \ll \mu$, a small ratio π_θ/μ reduces the gradient contribution in Standard IS. APRO combines adaptive ratio switching with discrepancy-aware scaling to strengthen learning from external trajectories. Appendix C.3 explains the local gradient mechanism, and Table 3 studies the separate contributions of switching and scaling.

Table 7: Additional results on continued on-policy RL and mixed-policy RL. For each backbone, **bold** and underlined mark the best and second-best value in each column; ties share the same mark. "+RL" denotes that the checkpoint obtained by the corresponding first-stage method is further optimized with on-policy GRPO. APRO(Mix) and TBRM(Mix) each use four current-policy rollouts and four external trajectories per prompt. These results evaluate whether off-policy RL serves as a useful initialization stage for subsequent self-exploration.

Method	In-Distribution Performance							Out-of-Distribution Performance		
	AIME 24	AIME 25	AMC	MATH-500	Minerva	Olympiad	Avg.	ARC-c	MMLU-Pro	Avg.
Qwen-2.5-Math-1.5B										
Base	7.4	3.0	23.3	34.0	9.6	18.7	16.0	2.6	4.2	3.4
SFT	11.1	11.4	36.2	68.6	22.8	36.1	31.0	36.2	19.1	27.7
RWR	12.8	13.9	37.9	69.4	27.6	34.2	32.6	32.2	20.6	26.4
AWR	11.6	14.1	37.9	72.0	27.2	34.2	32.8	25.5	18.8	22.2
NLL-UL	11.8	13.6	36.8	70.4	28.3	33.3	32.4	30.7	20.1	25.4
APRO(GRPO)	12.3	10.3	39.5	71.2	27.9	37.7	33.2	40.5	22.3	31.4
SFT + RL	**15.8**	**14.6**	42.5	72.4	27.9	37.7	35.2	38.4	22.3	30.4
RWR + RL	13.0	12.0	42.7	**77.4**	28.3	40.0	35.6	32.6	20.9	26.8
AWR + RL	14.7	14.4	**44.2**	73.8	26.5	38.8	35.4	38.7	23.6	31.2
NLL-UL + RL	13.5	12.9	43.0	75.4	29.4	41.3	35.9	38.9	23.2	31.1
TBRM (Mix)	11.3	12.1	35.6	71.2	27.5	29.4	31.2	29.2	23.1	26.2
TBRM + RL	12.1	12.7	36.5	69.2	28.1	30.2	31.5	32.5	24.7	28.6
APRO(Mix)	15.0	13.1	42.7	74.4	30.5	41.5	36.2	46.1	28.8	37.5
APRO(GRPO) + RL	14.1	14.2	42.4	74.6	**30.9**	**42.2**	**36.4**	**49.3**	**29.6**	**39.5**
DeepSeek-R1-Distill-Qwen-1.5B										
Base	19.0	16.7	51.2	72.8	23.5	40.0	37.2	5.7	18.4	12.1
SFT	20.5	17.4	55.1	81.4	30.9	45.2	41.8	18.8	19.5	19.2
RWR	23.0	20.8	54.7	79.4	27.9	43.5	41.6	11.0	14.9	13.0
AWR	24.3	19.4	54.6	77.8	31.3	43.8	41.9	10.8	14.1	12.5
NLL-UL	21.6	19.5	52.7	80.8	30.1	43.4	41.4	11.5	12.9	12.2
APRO(GRPO)	23.1	22.2	59.9	84.0	30.9	51.6	45.3	36.6	29.5	33.1
SFT + RL	27.2	20.2	59.3	81.8	**34.9**	49.8	45.5	23.8	23.0	23.4
RWR + RL	25.0	21.3	58.3	82.2	32.7	49.6	44.9	26.7	20.6	23.7
AWR + RL	24.8	**23.1**	58.4	83.2	34.2	49.6	45.6	33.0	22.2	27.6
NLL-UL + RL	**27.8**	21.0	56.1	82.8	33.1	47.3	44.7	20.1	18.1	19.1
TBRM (Mix)	24.6	21.4	60.4	82.0	30.5	51.4	45.1	22.2	16.4	19.3
TBRM + RL	24.8	21.8	60.2	82.6	31.3	50.8	45.3	26.7	23.9	25.3
APRO(Mix)	25.0	21.9	61.0	**84.2**	31.6	51.6	45.9	17.8	23.2	20.5
APRO(GRPO) + RL	25.5	23.0	**61.9**	83.6	33.5	**51.9**	**46.6**	**40.0**	**31.8**	**35.9**

H Additional Results: Off-Policy RL as a Cold-Start Stage

A natural question is whether fully off-policy RL is useful beyond the static off-policy setting. In practical post-training pipelines, external trajectories are often used to initialize a model, after which the policy may continue to improve through on-policy exploration. From this perspective, fully off-policy RL should not be viewed only as a standalone alternative to on-policy RL, but also as a capability-transfer stage that prepares a weak target policy for more effective subsequent exploration. This is particularly relevant for small models: when the initial policy rarely samples correct reasoning traces, purely on-policy RL can be limited by sparse rewards, while SFT-style imitation can provide a useful initialization but does not directly perform reward-sensitive policy improvement over positive and negative trajectories.

To examine this issue, we conduct two additional experiments. First, we evaluate *continued on-policy RL*, denoted by "+RL". For each method, we first train the model with the corresponding offline objective, including SFT, RWR, AWR, NLL-UL, TBRM, or APRO, and then use the resulting checkpoint as the initialization for on-policy GRPO under the same reward function. This setting tests whether the first-stage objective produces a better starting policy for later self-exploration. Second, we evaluate mixed-policy variants of APRO and TBRM, denoted by APRO(Mix) and TBRM(Mix). In this setting, each prompt is associated with eight trajectories: four trajectories are sampled from the current policy, and the remaining four are external off-policy trajectories. This setup directly tests whether external trajectories remain useful when on-policy rollouts are also available. The DeepSeek comparisons use matched first-stage updates, trajectory counts and batch sizes, followed by the same on-policy GRPO protocol.

Table 7 further supports the role of off-policy RL as a cold-start stage rather than an isolated training regime. Although SFT and regression-style objectives provide strong initial checkpoints, their gains after subsequent on-policy GRPO are relatively limited, especially on OOD benchmarks. On

Qwen-2.5-Math-1.5B, SFT+RL reaches 35.2 ID average and 30.4 OOD average, while the best imitation-style +RL baseline remains close in ID performance but does not close the OOD gap. In contrast, APRO(GRPO)+RL reaches 36.4 ID average and 39.5 OOD average. TBRM+RL reaches 31.5 ID and 28.6 OOD on this backbone. The large OOD improvement suggests that APRO does not simply fit the static external trajectories; it produces a policy that is better positioned for later self-exploration.

The same pattern appears on DeepSeek-R1-Distill-Qwen-1.5B. SFT+RL obtains a strong ID average of 45.5, showing that imitation is an effective cold-start strategy. However, its OOD average remains 23.4, and the best imitation-style +RL baseline reaches only 27.6. APRO(GRPO)+RL improves the OOD average to 35.9 while also achieving the best ID average of 46.6. TBRM+RL reaches 45.3 ID and 25.3 OOD. This indicates that reward-driven off-policy optimization can transfer external behaviors in a way that is more compatible with subsequent on-policy policy improvement than likelihood-based imitation alone.

We also evaluate APRO(Mix), where each prompt contains both on-policy rollouts and external off-policy trajectories. This setting is closer to practical post-training pipelines, where self-generated samples and external traces may coexist. APRO(Mix) performs strongly on Qwen-2.5-Math-1.5B, reaching 36.2 ID average and improving the OOD average over its static counterpart, showing that external trajectories remain useful even when on-policy rollouts are available. On Qwen, APRO(Mix) reaches 36.2 ID and 37.5 OOD versus 31.2 and 26.2 for TBRM(Mix). On DeepSeek, the corresponding APRO/TBRM averages are 45.9/45.1 ID and 20.5/19.3 OOD. Both methods obtain higher OOD scores after continued GRPO than in the mixed-policy setting, showing the value of examining these training schedules separately.

Overall, these results clarify the intended role of fully off-policy RL. It is not meant to replace on-policy self-exploration once the policy can generate useful rollouts. Instead, APRO serves as a capability-transfer stage: it uses external trajectories to move a weak policy into a better region of the policy space, after which on-policy GRPO can further amplify the gains. This distinguishes APRO from SFT, RWR, AWR, and NLL-UL, which mainly transfer demonstrations through likelihood or regression-style objectives, but do not explicitly perform behavior-corrected policy improvement over reward-labeled off-policy trajectories.

I Additional Diagnostics

The following diagnostics use DeepSeek-R1-Distill-Qwen-1.5B for mathematical reasoning.

I.1 Gradient-Norm Statistics

Figure 3 displays gradient norms on a logarithmic axis. To describe the APRO curve more precisely, Table 8 groups the gradient norms into six consecutive training intervals. The means rise early and stabilize near 3.0 in the final two intervals. The largest observed value, 4.758, occurs at step 175.

Table 8: Gradient-norm statistics for APRO in Figure 3.

Training steps	Mean	Std.	Median	Maximum
1–50	0.149	0.050	0.132	0.406
51–100	0.330	0.082	0.316	0.619
101–150	1.432	0.788	1.475	2.720
151–200	2.639	0.649	2.489	4.758
201–250	3.000	0.592	3.062	4.096
251–288	3.037	0.559	3.005	4.491

I.2 Projected Behavior Policy Sensitivity

We vary the training of the projected behavior policy while retaining the APRO optimization procedure. Table 9 compares checkpoints trained on all trajectories for one, three, and five epochs with projectors trained on positive-only or negative-only trajectories. The five-epoch all-trajectory projector has the highest OOD average, whereas the one-epoch projector has the highest ID average. Projectors fit to only one reward subset have lower ID and OOD averages than the five-epoch all-trajectory setting.

Table 9: Sensitivity to projected behavior policy training on mathematical reasoning. "All" includes positive and negative trajectories.

Projector training	ID Avg.	OOD Avg.
All trajectories, 1 epoch	45.8	26.5
All trajectories, 3 epochs	44.7	29.7
All trajectories, 5 epochs	45.3	33.1
Positive only	43.4	23.9
Negative only	43.7	24.1