

ACTOR-CURATOR: SCALABLE SELF-DRIVEN CURRICULUM LEARNING FOR LLM POST-TRAINING

Anonymous authors

Paper under double-blind review

ABSTRACT

Post-training large foundation models with reinforcement learning typically involves selecting training problems from massive and heterogeneous datasets, where the choice of data has a critical impact on training stability, sample efficiency, and final performance. In this work, we propose ACTOR-CURATOR, a scalable and fully automated framework for reinforcement learning post-training of large language models (LLMs) that learns to adaptively curate training problems. ACTOR-CURATOR trains a neural *curator* that dynamically selects problems from large problem banks by directly optimizing for expected policy performance improvement. We formulate problem selection as a non-stationary stochastic bandit problem, derive a principled loss function based on online stochastic mirror descent, and establish regret guarantees under partial feedback. Empirically, ACTOR-CURATOR consistently outperforms uniform sampling and strong learning-based baselines across a wide range of challenging reasoning benchmarks, demonstrating improved training stability and efficiency. Notably, it achieves relative gains of **28.6% on AIME2024** and **30.5% on ARC-1D** over the strongest baseline and up to **80% speedup**. These results suggest that ACTOR-CURATOR provides a practical and principled approach to scalable, adaptive curriculum learning for LLM post-training.

1 INTRODUCTION

Reinforcement learning (RL) has become a central paradigm for post-training foundation models, enabling improvements in reasoning, alignment, and task-specific performance beyond supervised fine-tuning (Shao et al., 2024). In this setting, the choice, ordering, and frequency of training problems play a critical role in determining convergence speed, training stability, and final generalization performance, motivating the use of curriculum learning to adaptively select training data (Bengio et al., 2009; Tzannetos et al., 2023; Parashar et al., 2025). However, applying curriculum learning to modern foundation model post-training is challenging: post-training datasets are **large, diverse, and continuously evolving**, while actor updates induce **complex, non-stationary training dynamics**. Traditional curriculum learning approaches—based on manual difficulty annotations, hand-designed problem buckets, or tabular per-problem statistics (Asada et al., 1996; Wu & Tian, 2017; Yengera et al., 2021)—do not scale to such settings, fail to generalize to unseen problems, and are brittle when problem utility changes as the policy improves. More-

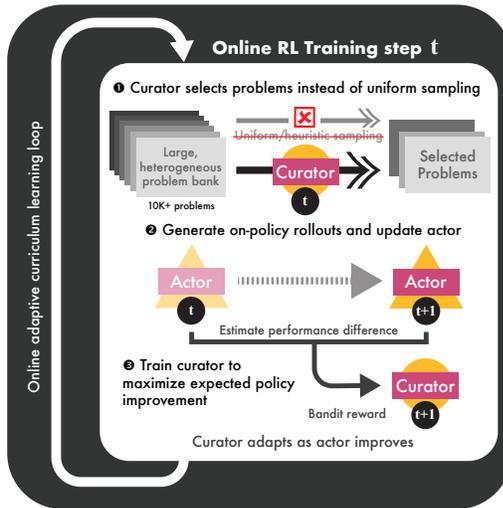


Figure 1: **Online training loop of ACTOR-CURATOR.** At each reinforcement learning step, a learned curator adaptively selects problems from a large problem bank instead of uniform sampling. The actor is updated on these problems, after which a bandit-style reward based on post-update policy improvement is used to train the curator. As the actor improves, the curator adapts to prioritize problems that yield the greatest expected performance gains.

over, effective curricula must balance **exploration** of under-sampled problems with **exploitation** of those that most improve the current policy, further complicating scalable curriculum design.

In this work, we propose ACTOR-CURATOR (AC), a scalable and fully automated problem curation framework for RL post-training that jointly trains an actor and a curator in an online, on-policy manner. At the core of ACTOR-CURATOR is a learned *curator* that adaptively selects training problems at each iteration and function-approximates over large, heterogeneous datasets. The curator is trained to directly maximize a **policy improvement objective**, assigning higher probability to problems expected to induce the greatest improvement in the actor’s performance. Unlike prior curricula that rely on heuristic signals such as absolute mean advantage or difficulty proxies (Chen et al., 2025a; Gao et al., 2025; Wang et al., 2025), our objective is derived from expected policy improvement, providing a principled and actor-aware learning signal. As a result, ACTOR-CURATOR naturally *adapts to evolving actor training dynamics* and allows it to be seamlessly combined with a wide range of RL algorithms.

To optimize the curator, we formalize problem selection as a **non-stationary stochastic bandit** problem with partial feedback. The curator is trained online and on-policy alongside the actor using **online stochastic mirror descent** (OSMD) Lattimore & Szepesvari (2017), which explicitly balances exploration and exploitation under non-stationarity. This differs from prior approaches that primarily rely on regression-style objectives and do not explicitly model the bandit structure or **partial observability** inherent in adaptive data selection (Tzannetos et al., 2023; Gao et al., 2025). To scale beyond tabular formulations, we derive a function-approximation variant of OSMD that trains the curator as a neural network, enabling generalization across problems and robustness to large, dynamic datasets. Finally, we introduce a PPO-style proximal clipping objective to stabilize curator optimization in practice.

Empirically, ACTOR-CURATOR enables effective curriculum learning at scale *without human annotations, difficulty labels, or manual dataset structuring*. Across diverse reasoning benchmarks—**Countdown, Zebra, MATH, AIME, and ARC-1D**—it consistently outperforms uniform sampling and strong baselines, achieving up to **30%** higher peak performance on ARC-1D, **28%** on AIME24, and up to **80%** faster convergence to comparable performance.

In summary, our main contributions are:

- **Automated problem curation for RL post-training.** We introduce ACTOR-CURATOR, a scalable framework that learns a neural curator to adaptively select training problems in an online, on-policy manner, enabling curriculum learning over large, heterogeneous datasets without human annotations or manual structuring.
- **A policy-improvement-driven bandit formulation of data curation.** We cast problem selection as a non-stationary stochastic bandit problem and derive a principled learning signal grounded in policy improvement theory, optimized via an OSMD-based bandit objective with regret guarantees under partial feedback.

2 PROBLEM FORMULATION

We study problem curation for reinforcement learning (RL) post-training of large language models (LLMs), where training is performed over large and heterogeneous collections of problems. Our goal is to design an adaptive data selection strategy that determines which training problems an LLM should train on at each iteration in order to maximize overall post-training performance.

2.1 RL POST-TRAINING SETTING

Let $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{|\mathcal{X}|}$ denote a large collection of training problems, and let $p_{\mathcal{X}}$ be a fixed evaluation distribution over \mathcal{X} . Let π denote a pretrained autoregressive language model, which induces a conditional distribution $\mathbf{y} \sim \pi(\cdot | \mathbf{x})$ over solutions for each problem \mathbf{x} . A reward model $R : \mathcal{Y} \times \mathcal{X} \rightarrow [0, 1]$ assigns a scalar score to each solution. The post-training objective is to maximize expected reward under the evaluation distribution:

$$J(\pi) \triangleq \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{X}}, \mathbf{y} \sim \pi(\cdot | \mathbf{x})} [R(\mathbf{y} | \mathbf{x})]. \quad (1)$$

In this work, the reinforcement learning algorithm, reward model, and rollout procedure are fixed. Our focus is on how training problems are selected across iterations.

2.2 TRAINING DYNAMICS

Training proceeds in iterations. At iteration t , a *curator* selects a subset of training problems $\mathcal{X}^t \subset \mathcal{X}$. Given this selection, the *actor* π^t is rolled out on each $\mathbf{x} \in \mathcal{X}^t$ to produce solutions and corresponding rewards

$$\mathcal{Y}_{\mathbf{x}}^t \triangleq \{\mathbf{y}^{(j)} \sim \pi^t(\cdot | \mathbf{x})\}_{j=1}^{|\mathcal{Y}_{\mathbf{x}}^t|}, \quad \mathcal{R}_{\mathbf{x}}^t \triangleq \{R(\mathbf{y}^{(j)} | \mathbf{x})\}_{j=1}^{|\mathcal{Y}_{\mathbf{x}}^t|}.$$

These trajectories form the dataset

$$\mathcal{D}^t \triangleq \{(\mathbf{x}, \mathcal{Y}_{\mathbf{x}}^t, \mathcal{R}_{\mathbf{x}}^t) | \mathbf{x} \in \mathcal{X}^t\}, \quad \pi^{t+1} \leftarrow \mathcal{A}(\pi^t, \mathcal{D}^t). \quad (2)$$

Here \mathcal{A} may correspond to any standard post-training algorithm (e.g., GRPO (Shao et al., 2024) or GSPO (Ahmadian et al., 2024)). We emphasize that the curator influences learning only indirectly through data selection, while the actor is solely responsible for policy optimization.

2.3 CURRICULUM LEARNING AS PROBLEM SELECTION

The central problem addressed in this work is how to choose the training subsets \mathcal{X}^t across iterations. Different choices of \mathcal{X}^t induce different actor updates and therefore different trajectories of policy improvement. We formalize curriculum learning as a sequential decision-making problem. At each iteration t , a *curator* selects a subset of training problems $\mathcal{X}^t \subset \mathcal{X}$. This selection induces a performance improvement $J(\pi^{t+1}) - J(\pi^t)$. The curator’s objective is to maximize cumulative performance gains over training:

$$\max_{\{\mathcal{X}^t\}_{t=1}^T} \sum_{t=1}^T (J(\pi^{t+1}) - J(\pi^t)). \quad (3)$$

2.4 CHALLENGES OF ADAPTIVE PROBLEM SELECTION

Effective curation is challenging for several reasons:

- **Large action space.** The problem set \mathcal{X} is large and might change across time, making it infeasible to manually define curricula or track per-problem statistics.
- **Partial feedback.** At each iteration, feedback is observed only for the problems selected for training; the utility of unselected problems remains unknown.
- **Non-stationarity.** The usefulness of a problem depends on the current actor π^t and changes as the actor improves. It is also highly dependent on the actor update method.
- **Exploration–exploitation trade-off.** The curator must balance exploring under-sampled problems whose utility is uncertain with exploiting problems that are known to drive policy improvement.

These challenges motivate a curriculum learning approach that operates at scale, learns online from partial feedback, and explicitly accounts for the non-stationary relationship between training problems and policy improvement.

3 METHOD

We now present ACTOR-CURATOR, a curriculum learning framework that trains a learned *curator* to adaptively select training problems for RL post-training of large language models. The key idea is to treat problem selection as a non-stationary bandit problem and to train the curator to directly maximize *policy improvement*—the expected performance gain induced by each actor update—using online stochastic mirror descent (OSMD) under partial feedback.

3.1 OVERVIEW OF THE TRAINING LOOP

Training proceeds in iterations. At iteration t , the following steps are performed:

1. We sample a training subset $\mathcal{X}^t \subset \mathcal{X}$ based on probabilities produced by the curator.
2. The actor π^t is rolled out on \mathcal{X}^t to collect trajectories and updated to π^{t+1} using any RL update.
3. Compute a per-problem policy improvement estimate for problems in \mathcal{X}^t using π^t and π^{t+1} .
4. The curator is updated online using bandit feedback derived from these improvement estimates.

The curator is trained jointly with the actor in an on-policy manner, allowing the curriculum to adapt dynamically as the actor improves. Figure 2 illustrates this process. Pseudocode is provided in Algorithm 1.

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

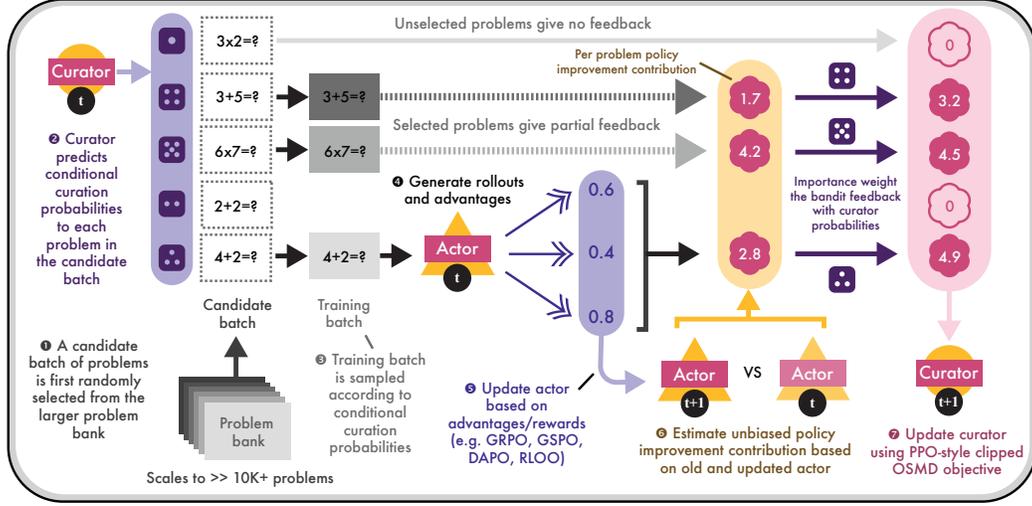


Figure 2: **Single training iteration of ACTOR-CURATOR.** At each iteration, a candidate set of problems is sampled from a fixed proposal distribution. The curator reweights this candidate set to select training problems for the actor. After the actor update, per-problem policy improvement is estimated using pre- and post-update policies. The curator observes bandit feedback only on selected problems and is updated using a PPO-style approximation of online stochastic mirror descent (OSMD).

3.2 POLICY IMPROVEMENT AS THE CURATOR LEARNING SIGNAL

The curator’s objective is to select problems that maximize improvement in the actor’s performance under the fixed evaluation distribution $p_{\mathcal{X}}$.

Performance improvement identity. Define the performance improvement at iteration t as

$$u^t \triangleq J(\pi^{t+1}) - J(\pi^t).$$

In the single-turn setting, the performance difference identity (Kakade & Langford, 2002) gives

$$u^t = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{X}}} \mathbb{E}_{\mathbf{y} \sim \pi^{t+1}(\cdot|\mathbf{x})} [A_{\pi^t}(\mathbf{y} | \mathbf{x})], \quad A_{\pi^t}(\mathbf{y} | \mathbf{x}) \triangleq R(\mathbf{y} | \mathbf{x}) - \mathbb{E}_{\mathbf{y}' \sim \pi^t(\cdot|\mathbf{x})} [R(\mathbf{y}' | \mathbf{x})]. \quad (4)$$

Applying importance sampling yields

$$u^t = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{X}}} \mathbb{E}_{\mathbf{y} \sim \pi^t(\cdot|\mathbf{x})} \left[\frac{\pi^{t+1}(\mathbf{y} | \mathbf{x})}{\pi^t(\mathbf{y} | \mathbf{x})} A_{\pi^t}(\mathbf{y} | \mathbf{x}) \right]. \quad (5)$$

Per-problem utility. Eq. (5) decomposes additively across problems. For each $\mathbf{x} \in \mathcal{X}$, define the per-problem utility

$$u_{\mathbf{x}}^t \triangleq p_{\mathcal{X}}(\mathbf{x}) \mathbb{E}_{\mathbf{y} \sim \pi^t(\cdot|\mathbf{x})} \left[\frac{\pi^{t+1}(\mathbf{y} | \mathbf{x})}{\pi^t(\mathbf{y} | \mathbf{x})} A_{\pi^t}(\mathbf{y} | \mathbf{x}) \right]. \quad (6)$$

By construction, $u^t = \sum_{\mathbf{x}} u_{\mathbf{x}}^t$. Although the actor update couples all selected problems, $u_{\mathbf{x}}^t$ provides a principled first-order credit assignment signal under small policy updates, grounded in policy improvement theory. We provide further explanation in App. H.

3.3 TABULAR OSMD FORMULATION

We cast curriculum learning as a non-stationary stochastic bandit problem, where each problem \mathbf{x} corresponds to an arm with time-varying utility $u_{\mathbf{x}}^t$. The curator is optimized using **online stochastic mirror descent** (OSMD) under bandit feedback (Lattimore & Szepesvari, 2017). We start with the tabular formulation first, where the curator maintains a probability mass function $\mathbf{p}^t \in \Delta_{\alpha}(\mathcal{X})$ over a finite set of problems, where \mathbf{p}^t is clipped to the sampling distribution $p^t(\mathbf{x} | \tilde{\mathcal{X}}^t) \geq \alpha > 0$. In Sec. 3.4 we show how to represent \mathbf{p}^t using a learned model.

Utility estimation and OSMD bandit feedback. For each $\mathbf{x} \in \mathcal{X}^t$, let $\mathcal{Y}_{\mathbf{x}}^t$ denote rollouts from $\pi^t(\cdot | \mathbf{x})$. We estimate Eq. (6) via

$$\hat{U}_{\mathbf{x}}^t \triangleq p_{\mathcal{X}}(\mathbf{x}) \frac{\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\}}{p^t(\mathbf{x})} \hat{A}^t(\cdot | \mathbf{x}), \quad \hat{A}^t(\cdot | \mathbf{x}) \triangleq \frac{1}{|\mathcal{Y}_{\mathbf{x}}^t|} \sum_{\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}^t} \frac{\pi^{t+1}(\mathbf{y} | \mathbf{x})}{\pi^t(\mathbf{y} | \mathbf{x})} A(\mathbf{y} | \mathbf{x}) \quad (7)$$

where \hat{A}^t is the importance normalized estimated average advantage. This estimate is agnostic to the specifics of the actor update method as long as an updated actor is produced.

Theorem 1 (Unbiasedness) $\mathbb{E}[\hat{U}_{\mathbf{x}}^t] = u_{\mathbf{x}}^t$.

We prove this in App. E. For decoder language models, $\pi^{t+1}(\mathbf{y} | \mathbf{x})$ is obtained via a single forward pass of the updated model on the previous solution.

OSMD update Given bandit feedback, the curator is updated with a negative-entropy regularizer:

$$p^{t+1} \leftarrow \arg \min_{\mathbf{p} \in \Delta_{\alpha}(\mathcal{X})} \left\{ -\eta \langle \mathbf{p}, \hat{U}^t \rangle + \text{KL}(\mathbf{p} \| \mathbf{p}^t) \right\}. \quad (8)$$

This yields the exponentiated-gradient update $p^{t+1}(\mathbf{x}) \propto p^t(\mathbf{x}) \exp(\eta \hat{U}_{\mathbf{x}}^t)$.

3.4 FUNCTION APPROXIMATION FOR CURATOR TRAINING

Although the OSMD update in Eq. (8) is defined over a distribution on the entire problem set \mathcal{X} , explicitly maintaining and updating tabular probabilities is infeasible when \mathcal{X} is large. We therefore parameterize the curator using a neural network that assigns a positive score to each problem and implicitly defines a probability distribution. The curator and induced distribution are defined as

$$C_{\phi} : \mathbf{x} \mapsto w_{\phi}(\mathbf{x}), \quad p_{\phi}(\mathbf{x}) \triangleq \frac{w_{\phi}(\mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{X}} w_{\phi}(\mathbf{x}')}, \quad w_{\phi}(\mathbf{x}) > 0.$$

OSMD surrogate objective. To implement the OSMD update Eq. (8) with function approximation, we optimize the following surrogate objective:

$$\mathcal{L}_{\text{cur}}(\phi) = \text{KL}(p_{\phi} \| p^t) - \eta \langle p_{\phi}, \hat{U}^t \rangle, \quad (9)$$

where p^t denotes the curator distribution from iteration t .

Scalable sampling and curator optimization. To scale ACTOR-CURATOR to large problem banks, we employ a practical training design that combines efficient sampling with stable curator optimization. First, rather than sampling directly from the curator distribution over the full problem set, we adopt a two-stage sampling procedure that decouples global coverage from adaptive curation. A candidate subset of problems is drawn from a fixed proposal distribution, after which the curator reweights this subset to select training problems. This design preserves unbiased utility estimation while keeping sampling computationally tractable. Full details of the sampling scheme, estimator corrections, and regret guarantees are provided in App. B.

Second, to stabilize curator learning under neural function approximation, we optimize the curator using a proximal objective inspired by PPO-style clipping. This prevents overly large distribution shifts between successive curator updates while retaining the behavior of the underlying mirror-descent formulation. The full derivation and objective are described in App. C.

4 EXPERIMENTAL RESULTS

Benchmarks. We evaluate ACTOR-CURATOR on five reasoning and mathematics benchmarks. **Countdown** is an arithmetic puzzle requiring the composition of integers and operations to reach a target value (Stojanovski et al., 2025). **Zebra** is a symbolic logic puzzle that requires finding assignments satisfying a set of constraints (Stojanovski et al., 2025). **ARC-1D** is the one-dimensional variant of the Abstraction and Reasoning Corpus, designed to test inductive generalization (Chollet, 2019; Xu et al., 2023). **MATH500** consists of competition-level mathematics problems (Hendrycks et al., 2021). **AIME24** contains problems from the 2024 American Invitational Mathematics Examination. We additionally consider hard subsets (**countdown-hard**, **zebra-hard**, **arc-hard**) for validation. We train on **30K** problems for Countdown, Zebra, and ARC-1D, and **12K** MATH problems for MATH500 and AIME24.

Table 1: **Peak validation performance** on problems within 100 training steps for different methods with qwen2.5-3b. ACTOR-CURATOR outperforms both other learning based methods (PCL) and methods that rely on human heuristics (SEC). We see similar results with llama3.2-3b-instruct, as shown in Tab. 4.

\mathcal{X}	BENCHMARK	METHOD					IMPROVEMENT	
		π_{ref}	UNIFORM	SEC	PCL	AC (OURS)	+ Δ	+ $\Delta\%$
30,000	COUNTDOWN	0.00	44.74	58.87	57.24	62.12	+3.25	+5.52
	COUNTDOWN-HARD	0.00	41.00	51.50	48.00	58.00	+6.50	+12.62
30,000	ZEBRA	0.00	35.12	36.00	34.12	37.62	+1.62	+4.50
	ZEBRA-HARD	0.00	30.50	27.50	26.00	34.50	+4.00	+13.11
30,000	ARC-1D	0.00	26.74	27.87	26.37	36.37	+8.50	+30.51
	ARC-HARD	0.00	19.50	18.50	18.50	31.00	+11.50	+58.97
12,000	MATH500	61.80	83.00	81.00	79.79	81.00	-2.00	-2.41
	AIME24	3.33	23.33	20.00	23.33	30.00	+6.67	+28.57

Experimental setup. We implement our post-training pipeline using VERL (Sheng et al., 2024). The curator is initialized from a pretrained Qwen3-0.6B model (Yang et al., 2025). Unless otherwise specified, the actor is trained using GSPO, a stabilized variant of GRPO (Ahmadian et al., 2024), on Qwen2.5-3B. Additional details are provided in App. J, including hyper-parameters.

4.1 MAIN RESULTS

Baselines. We compare against state-of-the-art curriculum learning methods for LLM post-training, using the same backbone model and actor update. π_{ref} denotes the pretrained model without post-training. **Uniform sampling** draws training problems uniformly at random. **SEC** partitions problems into manually defined buckets and updates bucket probabilities based on the sum of absolute advantages (Chen et al., 2025a). **PCL** trains a value model to estimate success probabilities and prioritizes problems with predicted success near 50% (Gao et al., 2025). PCL is competitive with recent curriculum-based approaches (Yue et al., 2025; Zhang et al., 2025; Zheng et al., 2025b).

Performance. We evaluate performance on held-out test sets, recording metrics every 10 training steps. Following prior work (Gao et al., 2025), we report the peak performance achieved within the first 100 steps. Results are summarized in Tab. 1. Across both backbone models and most benchmarks, ACTOR-CURATOR consistently outperforms all baselines, with additional results in App. M. Notably, ACTOR-CURATOR achieves substantially larger gains on harder benchmarks such as **arc-hard** and **AIME24**, indicating that adaptive curation is particularly beneficial in challenging regimes.

Efficiency. As shown in Fig. 4, ACTOR-CURATOR reaches comparable or higher performance using significantly fewer training steps than uniform sampling, demonstrating improved sample efficiency.

Training dynamics. Across datasets, ACTOR-CURATOR exhibits more stable optimization and often continues to improve performance after baselines plateau, as shown in Figs. 3, 9 and 10.

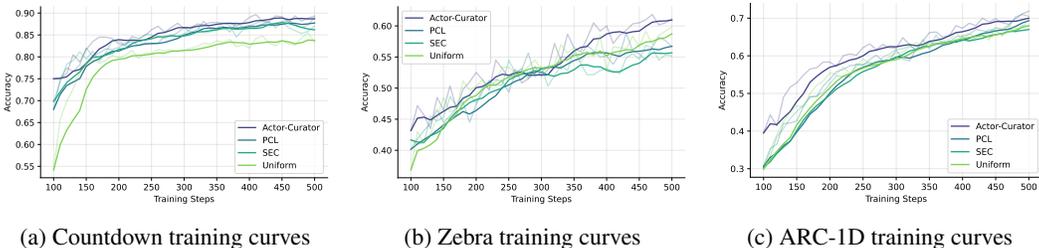


Figure 3: **Training dynamics.** ACTOR-CURATOR Test performance over training on three datasets, showing faster convergence and higher final accuracy.

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377

4.2 ABLATIONS

Core components. We ablate two key design choices: the policy-improvement utility and the OSMD bandit objective. **Absolute adv** replaces the policy-improvement signal with mean absolute advantage, as in (Chen et al., 2025a). **Regression loss** trains the curator to predict the target utility value using a squared loss, rather than learning a classifier as in OSMD. The predicted utilities are then converted into a sampling distribution via a Boltzmann transform with temperature η , matching the temperature used in OSMD. As shown in Fig. 7b, both components are critical for achieving strong performance. We include a more detailed account of the motivation behind Absolute Adv. and regression in App. D, as well as their difference with ACTOR-CURATOR.

Actor update methods. ACTOR-CURATOR generalizes across actor optimization algorithms. In Fig. 7a, ACTOR-CURATOR significantly improves GRPO-based training relative to uniform sampling, demonstrating robustness to the choice of actor update.

Additional ablation. We provide additional ablation on curator model size (Fig. 7c) and candidate batch size (Fig. 8) in App. L.

4.3 INTERPRETATION AND ANALYSIS

Curriculum progression. As shown in Fig. 5, ACTOR-CURATOR initially prioritizes easier problems and gradually shifts toward harder ones over training. Problem difficulty is estimated using heuristic annotations.

Impact on actor updates. ACTOR-CURATOR induces consistently larger actor gradient norms than uniform sampling (Fig. 6), suggesting that curated problems produce more informative learning signals. This aligns with the curator’s objective of prioritizing problems with higher expected policy improvement.

5 RELATED WORK

RLVR. Reinforcement learning with verification (RLVR) has emerged as an effective paradigm for improving the capabilities of large language models (LLMs) during post-training (Guo et al., 2025; Setlur et al., 2024; Chen et al., 2025b; Wen et al., 2025). Prior work has largely focused on improving the actor update rules (Yu et al., 2025; Dong et al., 2025) or enhancing trajectory generation, often via search-based methods (Zhang et al., 2024; Light et al., 2025b;c). Our work is complementary: rather than modifying the actor or rollout process, we focus on learning a principled curriculum that selects which problems the actor should train on to maximize policy improvement.

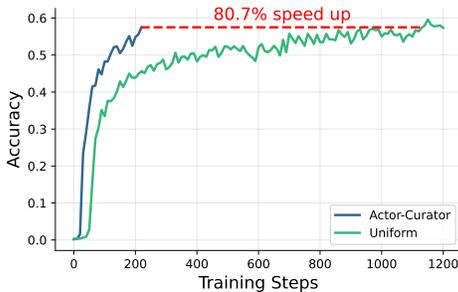


Figure 4: **Training speed-up.** ACTOR-CURATOR attains high test accuracy with significantly fewer steps.

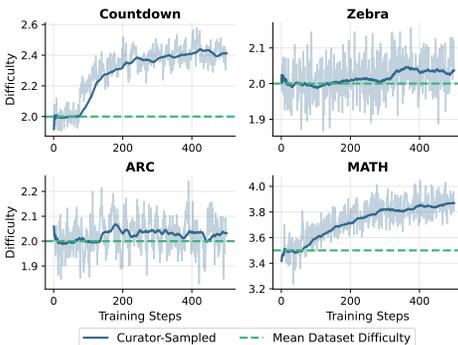


Figure 5: **Difficulty progression of curated problems.** ACTOR-CURATOR gradually increases the average difficulty over training.

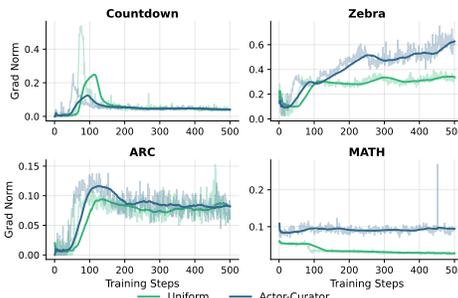


Figure 6: **Actor gradient norms over training.** ACTOR-CURATOR yields larger and more sustained updates.

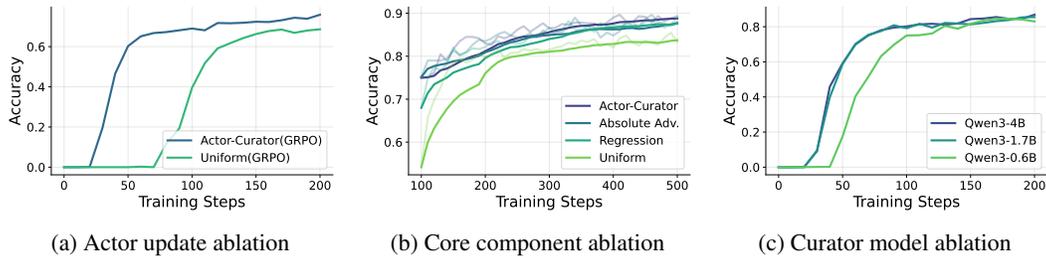


Figure 7: **Ablation results.** (a) ACTOR-CURATOR is compatible with alternative actor update methods such as GRPO and yields consistent performance gains. (b) The combination of the OSMD curator objective and the policy-improvement target achieves superior performance compared to alternative targets and loss functions. (c) Varying curator model size leads to similar long-term performance, indicating robustness to curator capacity.

Curriculum learning for RL. Classical curriculum learning approaches in reinforcement learning typically select tasks of intermediate difficulty, often defined via the agent’s probability of success (Florensa et al., 2017; 2018; Wöhlke et al., 2020; Liu et al., 2025). For example, ProCuRL selects problems whose difficulty is estimated to be near a decision boundary using a learned value network (Tzannetos et al., 2023). In contrast, our approach (1) trains a large language model as a curator that directly operates over language problems, and (2) uses a theoretically grounded, policy-improvement-based target rather than heuristic difficulty estimates. This design allows our method to generalize across different actor update rules and scale to large, heterogeneous datasets.

Curriculum learning for LLMs. Recent work on self-improving and self-evolving LLMs has highlighted the importance of curriculum learning in RL-based post-training (Ye et al., 2024; Light et al., 2025a). Several methods adjust problem sampling using bandit-style objectives such as UCB (Chen et al., 2025a; Wang et al., 2025; Gao et al., 2025). However, most existing approaches rely on manual curriculum design, including human-labeled difficulty levels or pre-defined problem buckets that are sampled adaptively (Graves et al., 2017). In contrast, we propose one of the first fully automated curriculum learning frameworks for LLM post-training that requires no human annotations or manual dataset structuring.

6 CONCLUSION

Our results demonstrate that combining bandit-style optimization, neural function approximation, and policy-improvement-based feedback provides a powerful and general approach to curriculum learning for RL post-training. This combination enables efficient and stable adaptation of training data selection at scale, leading to faster learning and higher final performance, particularly on difficult reasoning problems. Importantly, these gains persist across diverse benchmarks and settings, suggesting that this design generalizes beyond specific tasks or datasets and offers a scalable foundation for improving reinforcement learning post-training of large language models.

Beyond empirical gains, ACTOR-CURATOR highlights a broader shift in how post-training systems should be designed. Rather than treating training data as a static resource, our results suggest that data selection itself can be optimized online as part of the learning process, adapting in tandem with the evolving policy. This perspective opens the door to post-training pipelines that are less reliant on meticulous dataset engineering and more resilient to distributional mismatch, noise, and continual data growth.

REFERENCES

- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.
- Minoru Asada, Shoichi Noda, Sukoya Tawaratsumida, and Koh Hosoda. Purposive behavior acquisition for a real robot by vision-based reinforcement learning. *Machine learning*, 23(2):279–303, 1996.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.
- Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/91ba7292e5388b90b58d0b839a7f19ec-Paper.pdf.
- Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamaloo. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*, 2025a.
- Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*, 2025b.
- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025.
- Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*, pp. 482–495. PMLR, 2017.
- Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pp. 1515–1528. PMLR, 2018.
- Zhaolin Gao, Joongwon Kim, Wen Sun, Thorsten Joachims, Sid Wang, Richard Yuanzhe Pang, and Liang Tan. Prompt curriculum learning for efficient llm post-training. *arXiv preprint arXiv:2510.01135*, 2025.
- Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pp. 1311–1320. Pmlr, 2017.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.
- 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.
- Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.
- Tor Lattimore and Csaba Szepesvari. Bandit algorithms. 2017. URL <https://tor-lattimore.com/downloads/book/book.pdf>.

- Jonathan Light, Min Cai, Weiqin Chen, Guanzhi Wang, Xiushi Chen, Wei Cheng, Yisong Yue, and Ziniu Hu. Strategist: Self-improvement of llm decision making via bi-level tree search. In *The Thirteenth International Conference on Learning Representations*, 2025a.
- Jonathan Light, Wei Cheng, Benjamin Riviere, Wu Yue, Masafumi Oyamada, Mengdi Wang, Yisong Yue, Santiago Paternain, and Haifeng Chen. Disc: Dynamic decomposition improves llm inference scaling. In *Advances in Neural Information Processing Systems*, 2025b.
- Jonathan Light, Yue Wu, Yiyu Sun, Wenchao Yu, Yanchi Liu, Xujiang Zhao, Ziniu Hu, Haifeng Chen, and Wei Cheng. Sfs: Smarter code space search improves llm inference scaling. In *The Thirteenth International Conference on Learning Representations*, 2025c.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.
- Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning, 2025. URL <https://arxiv.org/abs/2506.06632>.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*, 2024.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- 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.
- Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2505.24760*, 2025.
- Georgios Tzannetos, Bárbara Gomes Ribeiro, Parameswaran Kamalaruban, and Adish Singla. Proximal curriculum for reinforcement learning agents. *arXiv preprint arXiv:2304.12877*, 2023.
- Zhenting Wang, Guofeng Cui, Yu-Jhe Li, Kun Wan, and Wentian Zhao. Dump: Automated distribution-level curriculum learning for rl-based llm post-training. *arXiv preprint arXiv:2504.09710*, 2025.
- Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Tanglifu Tanglifu, Xiaowei Lv, et al. Light-rl: Curriculum sft, dpo and rl for long cot from scratch and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 318–327, 2025.
- Jan Wöhlke, Felix Schmitt, and Herke van Hoof. A performance-based start state curriculum framework for reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1503–1511, 2020.
- Yuxin Wu and Yuandong Tian. Training agent for first-person shooter game with actor-critic curriculum learning. In *International Conference on Learning Representations*, 2017.

- 540 Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. Llms and the abstrac-
541 tion and reasoning corpus: Successes, failures, and the importance of object-based representations.
542 *arXiv preprint arXiv:2305.18354*, 2023.
- 543
- 544 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
545 Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*,
546 2025.
- 547 Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V Le, Qijun Tan, and
548 Yuan Liu. Scalable reinforcement post-training beyond static human prompts: Evolving alignment
549 via asymmetric self-play. *arXiv preprint arXiv:2411.00062*, 2024.
- 550
- 551 Gaurav Yengera, Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. Curriculum design for
552 teaching via demonstrations: theory and applications. *Advances in Neural Information Processing*
553 *Systems*, 34:10496–10509, 2021.
- 554 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian
555 Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at
556 scale. *arXiv preprint arXiv:2503.14476*, 2025.
- 557
- 558 Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiase Chen, Chengyi
559 Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for
560 advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.
- 561 Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm
562 self-training via process reward guided tree search. *Advances in Neural Information Processing*
563 *Systems*, 37:64735–64772, 2024.
- 564
- 565 Ruiqi Zhang, Daman Arora, Song Mei, and Andrea Zanette. Speed-rl: Faster training of reasoning
566 models via online curriculum learning. *arXiv preprint arXiv:2506.09016*, 2025.
- 567
- 568 Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang,
569 Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint*
arXiv:2507.18071, 2025a.
- 570
- 571 Haizhong Zheng, Yang Zhou, Brian R Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and
572 Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective
573 rollouts. *arXiv preprint arXiv:2506.02177*, 2025b.
- 574
- 575
- 576
- 577
- 578
- 579
- 580
- 581
- 582
- 583
- 584
- 585
- 586
- 587
- 588
- 589
- 590
- 591
- 592
- 593

A ALGORITHM PSEUDO-CODE

Algorithm 1 Actor–Curator: Self-driven curriculum learning

```

1: Input: dataset  $\mathcal{X}$ , pretrained LLM (actor)  $\pi^0$ , reward model  $R$ , curator model  $C_{\phi^0}$ , proposal
distribution  $\tilde{q}$ 
2: Hyperparameters: number of training steps  $T$ , rollouts per problem  $|\mathcal{Y}_x|$ , candidate batch size
 $|\tilde{\mathcal{X}}^t|$ , training batch size  $|\mathcal{X}^t|$ 
3: for  $t = 0$  to  $T - 1$  do ▷ Training steps
4:    $\mathcal{D}^t \leftarrow \emptyset$ 
5:   Proposal step: sample a candidate batch  $\tilde{\mathcal{X}}^t \subset \mathcal{X}$  according to  $\tilde{q}$ 
6:   Selection step: sample a training batch  $\mathcal{X}^t \subset \tilde{\mathcal{X}}^t$  using the curator  $C_{\phi^t}$ 
7:   for each problem  $\mathbf{x} \in \mathcal{X}^t$  do
8:     Roll out solutions  $\mathcal{Y}_x^t = \{\mathbf{y}^{(j)} \sim \pi^t(\cdot | \mathbf{x})\}_{j=1}^{|\mathcal{Y}_x^t|}$ 
9:     Compute rewards  $\mathcal{R}_x^t = \{R(\mathbf{y}^{(j)} | \mathbf{x})\}_{j=1}^{|\mathcal{Y}_x^t|}$ 
10:    Add  $(\mathbf{x}, \mathcal{Y}_x^t, \mathcal{R}_x^t)$  to  $\mathcal{D}^t$ 
11:   end for
12:   Actor update:  $\pi^{t+1} \leftarrow \mathcal{A}(\pi^t, \mathcal{D}^t)$ 
13:   Curator utilities: compute  $\hat{U}^t$  using equation 10 (and  $\pi^{t+1}$ )
14:   Curator update:  $\phi^{t+1} \leftarrow \arg \min_{\phi} \mathcal{L}_{\text{cur}}(\phi)$  using equation 12
15: end for

```

594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

B TWO-STAGE SAMPLING

Sampling directly from a curator distribution over the full problem set \mathcal{X} is computationally infeasible at scale. We therefore adopt a two-stage sampling scheme that separates *coverage* from *adaptive curation* while preserving unbiased utility estimation.

At iteration t , we first sample a candidate set $\tilde{\mathcal{X}}^t \subset \mathcal{X}$ from a fixed proposal distribution \tilde{q} . Let

$$q(\mathbf{x}) \triangleq \Pr_{\tilde{\mathcal{X}} \sim \tilde{q}}(\mathbf{x} \in \tilde{\mathcal{X}})$$

denote the induced marginal inclusion probability, and assume $q(\mathbf{x}) \geq q_{\min} > 0$ for all $\mathbf{x} \in \mathcal{X}$.

Conditioned on $\tilde{\mathcal{X}}^t$, the curator samples a training set $\mathcal{X}^t \subset \tilde{\mathcal{X}}^t$ according to the restricted distribution

$$p^t(\mathbf{x} \mid \tilde{\mathcal{X}}^t) \triangleq \frac{w^t(\mathbf{x})}{\sum_{\mathbf{x}' \in \tilde{\mathcal{X}}^t} w^t(\mathbf{x}')},$$

where $w^t(\mathbf{x}) > 0$ is the curator score at iteration t . This allows the curator to prioritize problems while operating only on a small candidate batch.

Utility estimation. The unbiased two-stage estimator corresponding to Eq. (7) is

$$\hat{U}_{\text{two}, \mathbf{x}}^t \triangleq p_{\mathcal{X}}(\mathbf{x}) \frac{\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\}}{q(\mathbf{x}) p^t(\mathbf{x} \mid \tilde{\mathcal{X}}^t)} \hat{A}^t(\cdot \mid \mathbf{x}), \quad (10)$$

which corrects for both proposal and curation sampling probabilities. This estimator satisfies $\mathbb{E}[\hat{U}_{\text{two}, \mathbf{x}}^t] = u_{\mathbf{x}}^t$ (see App. E). Substituting Eq. (10) into the OSMD update Eq. (8) yields the surrogate objective

$$\mathcal{L}_{\text{two}}(\phi) = \text{KL}(p_{\phi} \parallel p^t) - \eta \sum_{\mathbf{x} \in \mathcal{X}^t} \frac{p_{\phi}(\mathbf{x} \mid \tilde{\mathcal{X}}^t)}{p^t(\mathbf{x} \mid \tilde{\mathcal{X}}^t)} \frac{p_{\mathcal{X}}(\mathbf{x}) \hat{A}^t(\cdot \mid \mathbf{x})}{q(\mathbf{x})}, \quad (11)$$

where $p_{\phi}(\cdot \mid \tilde{\mathcal{X}}^t)$ is the conditional curator distribution.

Regret guarantee. We now state a regret bound for curator optimization under two-stage sampling. The bound characterizes the curator’s ability to track the best sequence of problem-selection distributions in hindsight despite non-stationary utilities. A proof is provided in App. F.

Theorem 2 *Assume the curator is updated using OSMD with a negative-entropy regularizer and receives bandit feedback $\hat{U}_{\text{two}, \mathbf{x}}^t$ satisfying $\mathbb{E}[\hat{U}_{\text{two}, \mathbf{x}}^t] = u_{\mathbf{x}}^t$. Then the cumulative dynamic regret satisfies*

$$\text{Reg}_T \leq O\left(T^{2/3} V_T^{1/3}\right), \quad V_T \triangleq \sum_{t=2}^T \max_{\mathbf{x} \in \mathcal{X}} |u_{\mathbf{x}}^t - u_{\mathbf{x}}^{t-1}|$$

where Reg_T is the regret against the best available arm, ignoring uniform exploration which we define in App. G, and V_T is a measure of how rapid the utility of a problem changes over time t .

While we focus on two-stage sampling for efficiency, ACTOR-CURATOR is compatible with other approximate sampling schemes (e.g., Metropolis–Hastings), provided marginal inclusion probabilities are roughly proportional to curator-assigned weights.

C PROXIMAL CURATOR OPTIMIZATION

Directly optimizing the KL-regularized objective in equation 9 can be unstable with neural network parameterization (Schulman et al., 2015). Following proximal policy optimization (PPO) (Schulman et al., 2017), we adopt a clipped surrogate objective. Define the importance ratio and sub-objective as

$$\rho_{\phi}(\mathbf{x}) \triangleq \frac{p_{\phi}(\mathbf{x} \mid \tilde{\mathcal{X}}^t)}{p^t(\mathbf{x} \mid \tilde{\mathcal{X}}^t)}, \quad g^t(\mathbf{x}) \triangleq \frac{p_{\mathcal{X}}(\mathbf{x}) \hat{A}^t(\cdot \mid \mathbf{x})}{q(\mathbf{x})}$$

Starting from equation 11, we replace the linear probability ratio term with a clipped surrogate

$$\mathcal{L}_{\text{cur}}^{\text{PCO}}(\phi) = -\eta \sum_{\mathbf{x} \in \tilde{\mathcal{X}}^t} \min\left(\rho_\phi(\mathbf{x}) g^t(\mathbf{x}), \text{clip}(\rho_\phi(\mathbf{x}), \rho_{\min}, \rho_{\max}) g^t(\mathbf{x})\right), \quad (12)$$

where ρ_{\min}, ρ_{\max} are clipping parameters. This objective constrains successive curator updates while preserving the behavior of online mirror descent in practice.

702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

D BASELINES

D.1 GROUP-BASED POLICY OPTIMIZATION

Group-based policy optimization methods update the policy by grouping multiple rollouts from the same problem and computing advantages relative to the group baseline.

GRPO. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) computes advantages by comparing each solution’s reward to the mean reward of all solutions sampled from the same problem. For a problem x with rollouts $\mathcal{Y}_x = \{y^{(1)}, \dots, y^{(m)}\}$ and corresponding rewards $\{R(y^{(j)}|x)\}_{j=1}^m$, the advantage for solution $y^{(i)}$ is:

$$A^{\text{GRPO}}(y^{(i)}|x) = R(y^{(i)}|x) - \frac{1}{m} \sum_{j=1}^m R(y^{(j)}|x). \quad (13)$$

The policy is then updated using a PPO-style clipped objective:

$$\mathcal{L}^{\text{GRPO}}(\pi) = \mathbb{E}_{x, y^{(i)} \sim \mathcal{Y}_x} \left[\min \left(\rho_i A^{\text{GRPO}}(y^{(i)}|x), \text{clip}(\rho_i, 1 - \epsilon, 1 + \epsilon) A^{\text{GRPO}}(y^{(i)}|x) \right) \right], \quad (14)$$

where $\rho_i = \frac{\pi(y^{(i)}|x)}{\pi_{\text{old}}(y^{(i)}|x)}$ is the importance ratio and ϵ is the clipping threshold.

GSPO. Group-Sequence Policy Optimization (Zheng et al., 2025a) addresses fundamental stability issues in GRPO by defining importance ratios at the sequence level rather than the token level. Unlike GRPO, which applies token-level importance weights that can introduce high-variance noise, GSPO computes importance ratios based on sequence likelihood, aligning with the principle of importance sampling.

GSPO optimizes the following sequence-level objective:

$$\mathcal{J}^{\text{GSPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^G \min \left(s_i(\theta) \hat{A}_i, \text{clip}(s_i(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_i \right) \right], \quad (15)$$

where the sequence-level importance ratio is defined as:

$$s_i(\theta) = \left(\frac{\pi_{\theta}(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)} \right)^{\frac{1}{|y_i|}} = \exp \left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log \frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} \right). \quad (16)$$

The advantage computation remains the same as in GRPO.

D.2 MEAN ABSOLUTE ADVANTAGE AS CURRICULUM REWARD

An ideal curriculum should prioritize training problems that maximize the model’s immediate learning outcomes. A natural way to quantify learning outcomes is through the magnitude of parameter updates induced by the selected training data, which can be approximated by the absolute advantage.

In the common setting of RL with verifiable binary rewards, mean absolute advantage has a attractive interpretation. When using group-based RL methods like GRPO with n rollouts per problem, the advantage for the i -th rollout is computed as:

$$\hat{A}_{t,i} = \frac{r_i - \text{mean}(r)}{\text{std}(r)} = \frac{r_i - p}{\sqrt{p(1-p)}}, \quad (17)$$

where p is the empirical success rate over the group. Since the reward r_i follows a Bernoulli distribution, the expected absolute advantage is:

$$\mathbb{E}[|\hat{A}_{t,i}|] = p \cdot \frac{1-p}{\sqrt{p(1-p)}} + (1-p) \cdot \frac{p}{\sqrt{p(1-p)}} = 2\sqrt{p(1-p)}. \quad (18)$$

The function $g(p) = 2\sqrt{p(1-p)}$ is symmetric around $p = 0.5$, strictly concave on $[0, 1]$, and reaches its maximum at $p = 0.5$. Thus, maximizing expected absolute advantage is equivalent to prioritizing problems at a success rate of 50%.

810 D.3 VALUE-BASED CURATION

811 Value-based curation methods maintain explicit estimates of the utility or learning value associated
812 with training data, which are then used to guide adaptive curriculum selection during RL post-training.

813 The core principle of value-based curation is to learn a utility function Q_t that maps training data
814 to expected learning outcomes at training step t . Let \mathcal{S} denote the space over which utilities are
815 estimated (e.g., individual problems, problem categories, or problem features). The utility function
816 $Q_t : \mathcal{S} \rightarrow \mathbb{R}$ assigns a scalar value to each element $s \in \mathcal{S}$.

817 During curriculum selection, the curator samples from \mathcal{S} according to a policy derived from Q_t . A
818 common choice is the Boltzmann (softmax) policy:
819

$$820 p_t(s) = \frac{\exp(Q_t(s)/\eta)}{\sum_{s' \in \mathcal{S}} \exp(Q_t(s')/\eta)}, \quad (19)$$

821 where $\eta > 0$ is a temperature parameter that controls the exploration-exploitation tradeoff: higher
822 temperatures lead to more uniform sampling (exploration), while lower temperatures concentrate
823 probability mass on high-utility items (exploitation).
824

825 The utility function is updated over time based on observed learning outcomes. After selecting
826 data according to p_t and performing an RL update, the curator observes a reward signal $r_t(s)$ that
827 measures the actual learning value obtained. These observations are used to refine Q_{t+1} .
828

829 **SEC.** Self-Evolving Curriculum (SEC) (Chen et al., 2025a) instantiates value-based curation by
830 defining the utility space $\mathcal{S} = \mathcal{C}$ as a discrete set of problem categories (e.g., difficulty levels, problem
831 types). The utility function $Q_t : \mathcal{C} \rightarrow \mathbb{R}$ is represented as a lookup table, with one scalar value per
832 category.
833

834 SEC uses temporal difference (TD) learning to update utilities:
835

$$836 Q_{t+1}(c) = \alpha r_t(c) + (1 - \alpha)Q_t(c), \quad (20)$$

837 where $\alpha \in (0, 1]$ is a learning rate and $r_t(c)$ is the mean absolute advantage aggregated over all
838 problems from category c selected at step t . This exponential moving average naturally adapts to
839 non-stationarity as category utilities change with model improvement.
840

841 Curriculum selection proceeds in two stages: first, a category is sampled according to the Boltzmann
842 policy $p_t(c) \propto \exp(Q_t(c)/\eta)$; then, problems are uniformly sampled from the selected category.
843

844 **Regression.** An alternative instantiation defines the utility space $\mathcal{S} = \mathcal{X}$ directly at the problem level,
845 estimating utilities for individual training problems. The utility function is parameterized by a neural
846 network $Q_\phi : \mathcal{X} \rightarrow \mathbb{R}$ that takes problem representations (e.g., text embeddings) as input.

847 Rather than incremental TD updates, the curator is trained via supervised regression. Let $\mathcal{H}_t =$
848 $\{(x_j, r_j)\}_{j=1}^t$ denote the history of problem-reward observations. The utility network is optimized to
849 minimize mean squared error:

$$850 \mathcal{L}_{\text{MSE}}(\phi) = \sum_{(x_j, r_j) \in \mathcal{H}_t} (Q_\phi(x_j) - r_j)^2. \quad (21)$$

851 This can be optimized periodically (every K steps) via batch gradient descent, or online via stochastic
852 gradient descent:
853

$$854 \phi_{t+1} = \phi_t - \beta \nabla_\phi (Q_\phi(x_t) - r_t(x_t))^2, \quad (22)$$

855 where β is the learning rate. Problems are sampled directly according to $p_t(x) \propto \exp(Q_\phi(x)/\eta)$.
856

857 **PCL** (Gao et al., 2025) exemplifies the online regression approach. PCL updates Q_ϕ concurrently
858 with policy training using only the current batch of observations. At each step t , PCL samples a
859 candidate pool of km prompts and selects the m prompts whose predicted values are closest to a
860 target threshold τ (typically 0.5):
861

$$862 D_m = \arg \min_{S \subseteq D_{km}, |S|=m} \sum_{x \in S} |Q_\phi(x) - \tau|. \quad (23)$$

864 This greedy selection can be viewed as an extreme limit of Boltzmann sampling with a sharply
 865 peaked distribution. Consider the modified utility $\tilde{Q}(x) = -|Q_\phi(x) - \tau|$, which measures negative
 866 distance from the threshold. As the temperature $\eta \rightarrow 0$, the Boltzmann policy $p(x) \propto \exp(\tilde{Q}(x)/\eta)$
 867 concentrates all probability mass on prompts nearest to τ , recovering PCL’s greedy selection. This
 868 deterministic selection strategy is computationally efficient and ensures the training batch contains
 869 only prompts of target difficulty, maximizing the effective ratio of informative gradient signals.

870 After selecting prompts, PCL generates n responses per prompt and updates both the policy and value
 871 model. The value model is trained on the observed rewards from the selected prompts:
 872

$$873 \mathcal{L}_{\text{PCL}}(\phi) = \sum_{i=1}^m \left(Q_\phi(x_i) - \frac{1}{n} \sum_{j=1}^n r(x_i, y_{i,j}) \right)^2. \quad (24)$$

877 **Comparison to Actor-Curator.** The Actor-Curator framework with OSMD differs fundamentally
 878 from value-based curation methods in its optimization objective and data selection mechanism.
 879 While value-based methods learn utilities $Q_t(s)$ to predict expected learning outcomes and sample
 880 accordingly, OSMD directly optimizes a curriculum distribution \mathbf{p}_t to maximize expected policy
 881 improvement. Value-based approaches require estimating problem-level or category-level values
 882 and making selection decisions based on these estimates—a two-stage process that introduces
 883 approximation error. In contrast, OSMD treats curriculum optimization as a first-order problem:
 884 the gradient $\nabla_q J(\pi_t, \mathbf{p})$ directly specifies how to adjust the data distribution to improve the policy,
 885 without requiring intermediate value estimates. Furthermore, value-based methods typically rely on
 886 scalar reward signals $r_t(s)$ to update utilities, whereas OSMD leverages the full gradient information
 887 $\nabla_\theta \mathcal{L}(\pi_t; x)$ to measure the learning value of each problem. This allows OSMD to capture richer
 888 information about how individual problems affect policy optimization, beyond what a single scalar
 889 reward can convey.

E POLICY IMPROVEMENT ESTIMATION

This section proves unbiasedness of the per-problem bandit feedback estimators used to train the curator. Recall the per-problem utility at iteration t (Eq. equation 6):

$$u_{\mathbf{x}}^t = p_{\mathcal{X}}(\mathbf{x}) \mathbb{E}_{\mathbf{y} \sim \pi^t(\cdot | \mathbf{x})} \left[\frac{\pi^{t+1}(\mathbf{y} | \mathbf{x})}{\pi^t(\mathbf{y} | \mathbf{x})} A_{\pi^t}(\mathbf{y} | \mathbf{x}) \right].$$

We also recall the rollout-based estimator (Eq. equation ??):

$$\hat{A}^t(\cdot | \mathbf{x}) \triangleq \frac{1}{|\mathcal{Y}_{\mathbf{x}}^t|} \sum_{\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}^t} \frac{\pi^{t+1}(\mathbf{y} | \mathbf{x})}{\pi^t(\mathbf{y} | \mathbf{x})} A_{\pi^t}(\mathbf{y} | \mathbf{x}), \quad \mathcal{Y}_{\mathbf{x}}^t = \{\mathbf{y}^{(j)} \sim \pi^t(\cdot | \mathbf{x})\}_{j=1}^{|\mathcal{Y}_{\mathbf{x}}^t|}.$$

E.1 SINGLE-STAGE CASE

Theorem 3 (Unbiasedness (single-stage)) *For the estimator in Eq. equation 7,*

$$\hat{U}_{\mathbf{x}}^t = p_{\mathcal{X}}(\mathbf{x}) \frac{\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\}}{p^t(\mathbf{x})} \hat{A}^t(\cdot | \mathbf{x}),$$

we have $\mathbb{E}[\hat{U}_{\mathbf{x}}^t] = u_{\mathbf{x}}^t$ for every $\mathbf{x} \in \mathcal{X}$.

Proof: Fix an iteration t and a problem $\mathbf{x} \in \mathcal{X}$. In the single-stage setting, the training set \mathcal{X}^t is sampled directly from \mathcal{X} according to the curator distribution $p^t(\cdot)$, so that

$$\Pr(\mathbf{x} \in \mathcal{X}^t) = p^t(\mathbf{x}).$$

First, $\hat{A}^t(\cdot | \mathbf{x})$ is an unbiased estimator of the population quantity

$$\mathbb{E}_{\mathbf{y} \sim \pi^t(\cdot | \mathbf{x})} \left[\frac{\pi^{t+1}(\mathbf{y} | \mathbf{x})}{\pi^t(\mathbf{y} | \mathbf{x})} A_{\pi^t}(\mathbf{y} | \mathbf{x}) \right],$$

by i.i.d. rollout sampling and linearity of expectation.

Next, take expectation of $\hat{U}_{\mathbf{x}}^t$ conditioning on $\hat{A}^t(\cdot | \mathbf{x})$:

$$\mathbb{E}[\hat{U}_{\mathbf{x}}^t | \hat{A}^t(\cdot | \mathbf{x})] = p_{\mathcal{X}}(\mathbf{x}) \mathbb{E} \left[\frac{\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\}}{p^t(\mathbf{x})} \right] \hat{A}^t(\cdot | \mathbf{x}) = p_{\mathcal{X}}(\mathbf{x}) \hat{A}^t(\cdot | \mathbf{x}),$$

since $\mathbb{E}[\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\}] = \Pr(\mathbf{x} \in \mathcal{X}^t) = p^t(\mathbf{x})$.

Finally, taking expectation over rollout randomness yields

$$\mathbb{E}[\hat{U}_{\mathbf{x}}^t] = p_{\mathcal{X}}(\mathbf{x}) \mathbb{E}[\hat{A}^t(\cdot | \mathbf{x})] = u_{\mathbf{x}}^t,$$

which proves the claim. \square

E.2 TWO-STAGE CASE

Theorem 4 (Unbiasedness (two-stage)) *For the two-stage estimator in Eq. equation 10,*

$$\hat{U}_{\text{two}, \mathbf{x}}^t = p_{\mathcal{X}}(\mathbf{x}) \frac{\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\}}{q(\mathbf{x}) p^t(\mathbf{x} | \tilde{\mathcal{X}}^t)} \hat{A}^t(\cdot | \mathbf{x}),$$

we have $\mathbb{E}[\hat{U}_{\text{two}, \mathbf{x}}^t] = u_{\mathbf{x}}^t$ for every $\mathbf{x} \in \mathcal{X}$.

Proof: Fix an iteration t and a problem $\mathbf{x} \in \mathcal{X}$. By definition of the two-stage procedure (Section B), the candidate set $\tilde{\mathcal{X}}^t$ is sampled from \tilde{q} , inducing the marginal inclusion probability $q(\mathbf{x}) = \Pr(\mathbf{x} \in \tilde{\mathcal{X}}^t)$. Conditioned on $\tilde{\mathcal{X}}^t$, the curator selects the training set $\mathcal{X}^t \subset \tilde{\mathcal{X}}^t$ according to $p^t(\cdot | \tilde{\mathcal{X}}^t)$.

As in the single-stage case, $\hat{A}^t(\cdot | \mathbf{x})$ is an unbiased estimator of the corresponding population expectation under $\mathbf{y} \sim \pi^t(\cdot | \mathbf{x})$.

972 Now condition on the realized candidate set $\tilde{\mathcal{X}}^t$ and on $\hat{A}^t(\cdot | \mathbf{x})$. If $\mathbf{x} \notin \tilde{\mathcal{X}}^t$, then $\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\} = 0$
 973 almost surely. If $\mathbf{x} \in \tilde{\mathcal{X}}^t$, then by the selection step,
 974

$$975 \mathbb{E}\left[\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\} \mid \tilde{\mathcal{X}}^t\right] = p^t(\mathbf{x} \mid \tilde{\mathcal{X}}^t),$$

976 and therefore

$$977 \mathbb{E}\left[\frac{\mathbb{I}\{\mathbf{x} \in \mathcal{X}^t\}}{p^t(\mathbf{x} \mid \tilde{\mathcal{X}}^t)} \mid \tilde{\mathcal{X}}^t, \hat{A}^t(\cdot | \mathbf{x})\right] = \mathbb{I}\{\mathbf{x} \in \tilde{\mathcal{X}}^t\}.$$

978 Substituting into the estimator gives

$$979 \mathbb{E}\left[\hat{U}_{\text{two}, \mathbf{x}}^t \mid \tilde{\mathcal{X}}^t, \hat{A}^t(\cdot | \mathbf{x})\right] = p_{\mathcal{X}}(\mathbf{x}) \frac{\mathbb{I}\{\mathbf{x} \in \tilde{\mathcal{X}}^t\}}{q(\mathbf{x})} \hat{A}^t(\cdot | \mathbf{x}).$$

980 Taking expectation over the proposal step yields

$$981 \mathbb{E}\left[\hat{U}_{\text{two}, \mathbf{x}}^t \mid \hat{A}^t(\cdot | \mathbf{x})\right] = p_{\mathcal{X}}(\mathbf{x}) \frac{\mathbb{E}[\mathbb{I}\{\mathbf{x} \in \tilde{\mathcal{X}}^t\}]}{q(\mathbf{x})} \hat{A}^t(\cdot | \mathbf{x}) = p_{\mathcal{X}}(\mathbf{x}) \hat{A}^t(\cdot | \mathbf{x}),$$

982 since $\mathbb{E}[\mathbb{I}\{\mathbf{x} \in \tilde{\mathcal{X}}^t\}] = \Pr(\mathbf{x} \in \tilde{\mathcal{X}}^t) = q(\mathbf{x})$. Finally, taking expectation over rollout randomness
 983 gives

$$984 \mathbb{E}[\hat{U}_{\text{two}, \mathbf{x}}^t] = p_{\mathcal{X}}(\mathbf{x}) \mathbb{E}[\hat{A}^t(\cdot | \mathbf{x})] = u_{\mathbf{x}}^t,$$

985 which proves the claim. □

986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

F IDEALIZED OSMD ALGORITHM

Algorithm 2 Sleeping Online Mirror Descent

Require: Number of total arms K , number of available arms k each round, horizon T , step size $\eta > 0$, exploration parameter $\alpha \in (0, 1/k)$.

1: Initialize $\mathbf{p}_1 = (1/K, \dots, 1/K)$.

2: **for** $t = 1, 2, \dots, T$ **do**

3: Sample available subset $\tilde{\mathcal{X}}_t$. Compute

$$\mathbf{p}^t(i | \tilde{\mathcal{X}}_t) = \begin{cases} \frac{\mathbf{p}_{t,i}}{\sum_{j \in \tilde{\mathcal{X}}_t} \mathbf{p}_{t,j}} & \text{if } i \in \tilde{\mathcal{X}}_t \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

4: Sample $a_{t,1}, a_{t,2}, \dots, a_{t,s} \stackrel{i.i.d.}{\sim} \mathbf{p}^t(\cdot | \tilde{\mathcal{X}}_t)$

5: (**Loss Estimator**) For each arm $i \in [K]$, set

$$\hat{L}_{t,i} = \frac{1}{s} \sum_{r=1}^s \mathbf{1}\{a_{t,r} = i\} \frac{l_{t,i}}{\mathbf{p}^t(i | \tilde{\mathcal{X}}_t)} \quad (26)$$

6: (**OSMD Update**) Update the next distribution by the mirror step

$$\mathbf{p}_{t+1} \in \arg \min_{\mathbf{p} \in \mathcal{A}} \left\{ \eta \langle \mathbf{p}, \hat{\mathbf{L}}_t \rangle + D_F(\mathbf{p}, \mathbf{p}_t) \right\}, \quad (27)$$

where $D_F(\mathbf{u}, \mathbf{v}) = \sum_i u_i \log \frac{u_i}{v_i}$

7: **end for**

F.1 SETUP

We formalize the tabular bandit algorithm in Section 3 as Algorithm 2. Under this idealized algorithm. We assume there is a large set of K arms. At each time step t , k arms are uniformly randomly chosen as candidate arms. The settings allows for the pulling of s arms per round, and after each round the loss $l_{t,i}$ is revealed for each chosen arm i , where i indicates the arm's original index in $[K]$. Throughout this section and the next, we denote the vectorized quantities with bold font. For example $\mathbf{l}_t = (l_{t,1}, \dots, l_{t,K})$.

Algorithm 2 departs from traditional Online Stochastic Mirror Descent (OSMD) due to the availability constraint. The action distribution is conditioned on a randomly sampled *available set* at each round, and the loss estimator is modified to remain unbiased under this conditional sampling.

Eqn. equation 25 introduces a two-stage sampling process. First, a random k -subset $\tilde{\mathcal{X}}_t \subseteq [K]$ of available arms is drawn uniformly. The learner then constructs a *conditional distribution*

$$\mathbf{p}^t(i | \tilde{\mathcal{X}}_t) = \frac{\mathbf{p}_{t,i}}{\sum_{j \in \tilde{\mathcal{X}}_t} \mathbf{p}_{t,j}} \quad \text{for } i \in \tilde{\mathcal{X}}_t,$$

and assigns zero probability to arms outside $\tilde{\mathcal{X}}_t$. We sometimes also use $\mathbf{p}_{|\tilde{\mathcal{X}}_t}(i)$ to denote $\mathbf{p}^t(i | \tilde{\mathcal{X}}_t)$. This renormalization ensures that the learner only samples from arms that are available at round t , while still using \mathbf{p}_t as the global state variable that is updated over time.

The conditional sampling in Eqn. equation 25 invalidates the standard OSMD estimator, since $\mathbf{p}_{t,i}$ is no longer the actual probability with which arm i is sampled. Eqn. equation 26 addresses this by defining

$$\hat{L}_{t,i} = \frac{1}{s} \sum_{r=1}^s \mathbf{1}\{a_{t,r} = i\} \frac{l_{t,i}}{\mathbf{p}^t(i | \tilde{\mathcal{X}}_t)}.$$

This estimator uses the conditional probability $\mathbf{p}^t(i | \tilde{\mathcal{X}}_t)$ in the denominator, which is the true sampling probability of arm i given the realized availability set. As a result, conditional on $\tilde{\mathcal{X}}_t$, the

estimator is unbiased:

$$\mathbb{E}[\widehat{L}_{t,i} \mid \tilde{\mathcal{X}}_t] = l_{t,i} \mathbb{I}\{i \in \tilde{\mathcal{X}}_t\}.$$

The use of s independent samples further reduces variance and corresponds to a semi-bandit feedback model, but does not change the role of the estimator in the mirror update.

G REGRET ANALYSIS

This section presents a proof of the regret bound (Theorem 2) for Algorithm 2. Suppose there are K base arms $[K] = \{1, \dots, K\}$ each corresponding to a problem in the dataset $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(K)}\}$. At each step, a subset $\tilde{\mathcal{X}}_t$ of size k is drawn uniformly randomly from \mathcal{X} . For every round t , The curator picks s arms $a_{t,1}, a_{t,2}, \dots, a_{t,s} \in [K]$ from $\tilde{\mathcal{X}}_t$ and the loss for each arm $l_{t,a_{t,i}}$ is revealed. For convenience, write $\mathbf{l}_t = (l_{t,1}, \dots, l_{t,K})$ to be the losses of each arm as a vector. Assume without loss of generality that $l_{t,i} \in [0, 1]$ for all t and i . Define the subset-masked loss vector $\mathbf{l}_t^{\tilde{\mathcal{X}}_t} \in \mathbb{R}^K$:

$$\mathbf{l}_t^{\tilde{\mathcal{X}}_t}(i) = \begin{cases} \mathbf{l}_t(i) & \text{if } i \in \tilde{\mathcal{X}}_t \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

Denote the best available arm at time t as m_t

$$m_t = \arg \min_{i \in \tilde{\mathcal{X}}_t} l_{t,i} \text{ and } l_t^* = l_{t,m_t}. \quad (29)$$

We define the *best-arm regret* to be

$$\text{Reg}_T^{\text{best}} = \mathbb{E} \left[\sum_{t=1}^T \left(\sum_{i=1}^s l_{t,a_{t,i}} - l_t^* \right) \right] \quad (30)$$

Note that the regret can be expressed in the vectorized form.

$$\text{Reg}_n^{\text{best}} = \mathbb{E} \left[\sum_{t=1}^T \langle \mathbf{p}_{\tilde{\mathcal{X}}_t}, \mathbf{l}_t \rangle - \langle \mathbf{e}_{m_t}, \mathbf{l}_t \rangle \right], \quad (31)$$

where \mathbf{e}_i is a one-hot vector with 1 on index i . Here, we use $\mathbf{p}_{|\tilde{\mathcal{X}}_t}$ to represent the vector $(\mathbf{p}_{|\tilde{\mathcal{X}}_t}(\mathbf{x}^{(1)}), \mathbf{p}_{|\tilde{\mathcal{X}}_t}(\mathbf{x}^{(2)}), \dots, \mathbf{p}_{|\tilde{\mathcal{X}}_t}(\mathbf{x}^{(K)}))$. In practice, forcing uniform exploration usually have negligible or even positive effect on the performance. However, it incurs linear regret for theoretical analysis. Thus, to focus on how OSMD manages losses in a non-stationary environment, this regret analysis isolate the loss attributable to factors other than uniform exploration. Towards this end, we set the comparator \mathbf{q}_t to be a mixture of the best available arm and uniform distribution over all arms.

$$\mathbf{q}_t = (1 - k\alpha)\mathbf{e}_{m_t} + \alpha \mathbf{1}_{\tilde{\mathcal{X}}_t} \quad (32)$$

where $\alpha > 0$ and $\mathbf{1}_{\tilde{\mathcal{X}}_t} \in \mathbb{R}^K$ is a binary vector with $\mathbf{1}_{\tilde{\mathcal{X}}_t}(i) = 1$ for all $i \in \tilde{\mathcal{X}}_t$ and zero everywhere else. The *best-available regret* is defined as

$$\text{Reg}_T^{\text{BA}} = \mathbb{E} \left[\sum_{t=1}^T \langle \mathbf{p}_{|\tilde{\mathcal{X}}_t}, \mathbf{l}_t \rangle - \langle \mathbf{q}_t, \mathbf{l}_t \rangle \right] \quad (33)$$

Theorem 5 (Restatement of Theorem 2) *Let the drift parameter*

$$V_T = \sum_{t=2}^T \Delta_t, \quad \Delta_t = \max_{i \in [K]} |l_{t,i} - l_{t-1,i}| \quad (34)$$

be a measure of how rapid the arm values changes over time t . Assume without loss of generality that $l_{t,i} \in [0, 1]$ for all t and i . Then we have

$$\text{Reg}_n^{\text{BA}} \leq O\left(T^{2/3} V_n^{1/3}\right) \quad (35)$$

for Algorithm 2.

To bound the smoothed regret, the proof follows a block decomposition argument. We partition the horizon into contiguous blocks of length B and restart the algorithm at the beginning of each block. Within a single block, the loss sequence is treated as approximately stationary, allowing us to compare the algorithm against a fixed comparator using standard OSMD analysis. The regret within each block is controlled by a stability–variance tradeoff: the mirror descent inequality yields a term of order $O(\log(1/\alpha)/\eta)$, while the variance of the importance-weighted estimator contributes a term proportional to $O(\eta B)$. Across blocks, non-stationarity is captured through a variation budget V_n , which upper bounds the cumulative discrepancy between the true per-round losses and the frozen losses used in the blockwise analysis, contributing an additive term of order BV_n . The block length B is then optimized to balance the statistical cost of restarting too frequently against the bias induced by treating losses as stationary within each block. This optimization yields a regret bound scaling as $O(V_n^{1/3})$, reflecting the classical tradeoff between adaptivity to non-stationarity and estimation error Besbes et al. (2014).

Let the length of each aforementioned block be L . Denote the start time of each block as $\tau_l = (l-1)L + 1$, where $l = 1, 2, \dots, B$ and $B = T/L^1$. Denote the time steps of the l -th block as $\mathcal{I}_l = \{\tau_l, \dots, \tau_l + L - 1\}$.

$$\text{Reg}_n^{\text{BA}} = \mathbb{E} \left[\sum_{t=1}^n \langle \mathbf{p}_{|\tilde{\mathcal{X}}_t}, \mathbf{l}_t \rangle - \langle \mathbf{q}_t, \mathbf{l}_t \rangle \right] = \sum_{l=1}^B \mathbb{E} \left[\sum_{t \in \mathcal{I}_l} \langle \mathbf{p}_{|\tilde{\mathcal{X}}_t}, \mathbf{y}_t \rangle - \langle \mathbf{q}_t, \mathbf{l}_t \rangle \right] \quad (36)$$

G.1 TIME-FROZEN COMPARATOR

This part of the proof introduces a stable reference for comparison in the presence of non-stationary losses. Rather than comparing the learner to the best action at every round, which may change arbitrarily over time, we define a comparator based on the losses at a fixed reference τ . This “time-frozen” comparator serves as a proxy for the per-round best action. The construction allows us to relate the learner’s loss to this frozen benchmark, while the error incurred by freezing time can be bounded by the amount of variation in the losses.

Define $m_t^\tau \in \arg \min_{i \in \tilde{\mathcal{X}}_t} y_{\tau, i}$ to be the τ -frozen best arm of time t . Denote the value associated with this arm $(y_t^*)^\tau = y_{\tau, m_t^\tau}$. Thus, we can define the time-frozen smooth comparator

$$\mathbf{q}_t^\tau = (1 - k\alpha) \mathbf{e}_{m_t^\tau} + \alpha \mathbf{1}_{\tilde{\mathcal{X}}_t} \quad (37)$$

Lemma 1 *Let $\mathcal{I} = \{\tau, \dots, \tau + L - 1\}$ be an arbitrary block. Suppose $\alpha > 0$. Let V_n be as defined in Theorem 5. For any sequence of random variables $\{X_t\}_{t \in \mathcal{I}}$ with a common support, we have*

$$\mathbb{E} \left[\sum_{t \in \mathcal{I}} X_t - \left\langle \mathbf{q}_t, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle \right] \leq \mathbb{E} \left[\sum_{t \in \mathcal{I}} X_t - \left\langle \mathbf{q}_t^\tau, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle \right] + \frac{1 - k\alpha}{k\alpha} \sum_{t \in \mathcal{I}} \|\mathbf{l}_t - \mathbf{l}_\tau\|_1 \quad (38)$$

Proof: By the definition of \mathbf{q}_t^τ , we have

$$\begin{aligned} \left\langle \mathbf{q}_t^\tau, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle &\leq (1 - k\alpha)(y_t^*)^\tau + \alpha \sum_{j \in \tilde{\mathcal{X}}_t} \mathbf{l}_{\tau, j} \\ &\leq (1 - k\alpha)(y_t^*)^\tau + \alpha(k(y_t^*)^\tau + k - 1) \\ &= m_t^\tau + \alpha(k - 1) \end{aligned} \quad (39)$$

Rearranging (39) and divide by Z_t , we get

$$\frac{m_t^\tau}{Z_t} \geq \left\langle \mathbf{q}_t^\tau, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle - \frac{\alpha(k-1)}{Z_t} \geq \left\langle \mathbf{q}_t^\tau, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle - \frac{k-1}{k}. \quad (40)$$

Using the fact that min is a Lipschitz operation, we have

$$|y_t^* - (y_t^*)^\tau| = \left| \min_{i \in \tilde{\mathcal{X}}_t} \mathbf{l}_{t, i} - \min_{i \in \tilde{\mathcal{X}}_\tau} \mathbf{l}_{\tau, i} \right| \leq \|\mathbf{l}_t - \mathbf{l}_\tau\|_\infty \leq \|\mathbf{l}_t - \mathbf{l}_\tau\|_1,$$

¹For convenience, we assume T is divisible by B and L at the same time.

Applying this identity, we have

$$\begin{aligned} \left| \left\langle \mathbf{q}_t, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle - \left\langle \mathbf{q}_t^\tau, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle \right| &\leq \frac{1}{k\alpha} \left| \left\langle \mathbf{q}_t, \mathbf{l}_t^{\tilde{\mathcal{X}}_t} \right\rangle - \left\langle \mathbf{q}_t^\tau, \mathbf{l}_t^{\tilde{\mathcal{X}}_t} \right\rangle \right| \\ &= \frac{1-k\alpha}{k\alpha} |y_t^* - (y_t^*)^\tau| \\ &\leq \frac{1-k\alpha}{k\alpha} \|\mathbf{y}_t - \mathbf{y}_\tau\|_1 \end{aligned} \quad (41)$$

Rearrange (41) and add X_t to both sides while taking expectation, we get

$$\mathbb{E} \left[\sum_{t \in \mathcal{I}} X_t - \left\langle \mathbf{q}_t, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle \right] \leq \mathbb{E} \left[\sum_{t \in \mathcal{I}} X_t - \left\langle \mathbf{q}_t^\tau, \frac{\mathbf{l}_t^{\tilde{\mathcal{X}}_t}}{Z_t} \right\rangle \right] + \frac{1-k\alpha}{k\alpha} \sum_{t \in \mathcal{I}} \|\mathbf{l}_t - \mathbf{l}_\tau\|_1 \quad (42)$$

□

G.2 MOMENT LEMMAS

To apply the standard OMD bound (Theorem 5), we require unbiased loss estimates with controlled variance. Since the learner only observes losses on the sampled subset ($\tilde{\mathcal{X}}_t$), we work with an importance-weighted estimator that accounts for both subset sampling and the normalization induced by Z_t .

The following two lemmas establish the properties needed for the regret analysis. The first shows that the estimator is unbiased for the scaled loss, ensuring that the expected update direction matches the true loss. The second provides a bound on the second moment of the estimator, which controls the variance term in the OSMD regret bound. Together, these results justify the use of the estimator in the mirror descent analysis and quantify the cost introduced by partial observation and uniform exploration.

Lemma 2 (Unbiasedness) Let $\hat{L}_{t,i}$ be as defined in Algorithm 2. Denote $\hat{\mathbf{L}}_t = (\hat{L}_{t,1}, \hat{L}_{t,2}, \dots, \hat{L}_{t,K})$. For every $i \in [K]$,

$$\mathbb{E}[\hat{L}_{t,i} \mid \tilde{\mathcal{X}}_t] = \mathbf{l}_{t,i} \mathbb{I}\{i \in \tilde{\mathcal{X}}_t\} = \mathbf{l}_t^{\tilde{\mathcal{X}}_t}(i) \quad (43)$$

Additionally, define the time-frozen estimator

$$\hat{L}_{t,i}^\tau := \frac{1}{s} \sum_{r=1}^s \mathbb{I}\{a_{t,r} = i\} \frac{l_{\tau,i}}{\mathbf{p}^\tau(i \mid \tilde{\mathcal{X}}_t)}, \quad (44)$$

and we have

$$\mathbb{E}[\hat{L}_{t,i}^\tau \mid \tilde{\mathcal{X}}_t] = l_{\tau,i} \mathbb{I}\{i \in \tilde{\mathcal{X}}_t\} = \mathbf{l}_\tau^{\tilde{\mathcal{X}}_t}(i) \quad (45)$$

Proof: If $i \in \tilde{\mathcal{X}}_t$,

$$\mathbb{E}[\hat{L}_{t,i}^\tau \mid \tilde{\mathcal{X}}_t] = \frac{1}{s} \sum_{r=1}^s \mathbf{p}_t(i \mid \tilde{\mathcal{X}}_t) \frac{l_{\tau,i} \mathbb{I}\{i \in \tilde{\mathcal{X}}_t\}}{\mathbf{p}^\tau(i \mid \tilde{\mathcal{X}}_t)} = l_{\tau,i} \mathbb{I}\{i \in \tilde{\mathcal{X}}_t\}$$

Similarly,

$$\mathbb{E}[\hat{L}_{t,i} \mid \tilde{\mathcal{X}}_t] = \frac{1}{s} \sum_{r=1}^s \mathbf{p}_t(i \mid \tilde{\mathcal{X}}_t) \frac{l_{t,i} \mathbb{I}\{i \in \tilde{\mathcal{X}}_t\}}{\mathbf{p}^t(i \mid \tilde{\mathcal{X}}_t)} = l_{t,i} \mathbb{I}\{i \in \tilde{\mathcal{X}}_t\}$$

□

Lemma 3 (Bounded Second Moment) Let $\hat{L}_{t,j}$ be as defined in Algorithm 2. Suppose we obtain \mathbf{p}^t through

$$\mathbf{p}^t \in \arg \min_{\mathbf{p} \in \mathcal{A}} \eta \left\langle \mathbf{p}, \hat{\mathbf{L}}_t \right\rangle + D_F(\mathbf{p}, \mathbf{p}^{t-1}),$$

under the constraint $\mathbf{p}^t(i) \geq \alpha$ for all i . We have

$$\mathbb{E} \left[\sum_{j=1}^K \mathbf{p}_j^t \hat{L}_{t,j}^2 \mid \tilde{\mathcal{X}}_t \right] \leq \frac{k}{s} \quad (46)$$

Proof: For $j \notin \tilde{\mathcal{X}}_t$, $\hat{L}_{t,j} = 0$, For $j \in \tilde{\mathcal{X}}_t$, let

$$X_{r,j} := \mathbb{I}\{A_{t,r=j}\} \frac{l_{r,j}}{\mathbf{p}^t(j | \tilde{\mathcal{X}}_t)}, \quad \text{consequently } \hat{L}_{t,j} = \frac{1}{s} \sum_{r=1}^s X_{r,j}$$

By construction, $X_{r,j}$'s are IID in r , and since $l_{r,j} \in [0, 1]$,

$$\mathbb{E} \left[X_{r,j}^2 | \tilde{\mathcal{X}}_t \right] = \sum_{j \in \tilde{\mathcal{X}}_t} \mathbf{p}^t(j | \tilde{\mathcal{X}}_t) \frac{l_{r,j}^2}{\mathbf{p}^t(j | \tilde{\mathcal{X}}_t)^2} \leq \frac{1}{\mathbf{p}^t(j | \tilde{\mathcal{X}}_t)}$$

By Jensen's inequality, we have

$$\mathbb{E} \left[(\hat{L}_{t,j}^\tau)^2 | \tilde{\mathcal{X}}_t \right] = \mathbb{E} \left[\left(\frac{1}{s} \sum_{r=1}^s X_{r,j} \right)^2 \middle| \tilde{\mathcal{X}}_t \right] \leq \frac{1}{s} \mathbb{E} \left[X_{r,j}^2 | \tilde{\mathcal{X}}_t \right] \leq \frac{1}{s \mathbf{p}^t(j | \tilde{\mathcal{X}}_t)}$$

Summing with weights \mathbf{p}^t over $j \in \tilde{\mathcal{X}}_t$ gives

$$\sum_{j \in \tilde{\mathcal{X}}_t} \left(\mathbf{p}_{|\tilde{\mathcal{X}}_t}^t \right)_j \mathbb{E} \left[(\hat{L}_{t,j}^\tau)^2 | \tilde{\mathcal{X}}_t \right] \leq \frac{1}{s} \sum_{j \in \tilde{\mathcal{X}}_t} \left(\mathbf{p}_{|\tilde{\mathcal{X}}_t}^t \right)_j \frac{Z_t}{\left(\mathbf{p}_{|\tilde{\mathcal{X}}_t}^t \right)_j} = \frac{k}{s} Z_t \leq \frac{k}{s}$$

□

G.3 ONE-STEP OMD BOUND

The purpose of the following lemma is to provide a local inequality that governs how a single mirror descent update behaves under the importance-weighted loss estimator. When summed over time, the KL terms telescope while the second-moment terms accumulate in a controlled manner. This structure allows the proof to separate the effect of the update rule from issues caused by partial observability and non-stationarity, which are handled in later steps.

Lemma 4 Suppose \mathbf{p}^t 's are obtained as in Algorithm 2. Then, for any comparator $\mathbf{v}_t \in \Delta_{\tilde{\mathcal{X}}_t}$,

$$\left\langle \mathbf{p}_{|\tilde{\mathcal{X}}_t}^t - \mathbf{v}_t, \mathbf{l}_t \right\rangle \leq \frac{1}{\eta} \left(\text{KL}(\mathbf{v}_t \| \mathbf{p}_{|\tilde{\mathcal{X}}_t}^t) - \text{KL}(\mathbf{v}_t \| \mathbf{p}_{|\tilde{\mathcal{X}}_{t+1}}^{t+1}) \right) + \frac{\eta s^2}{2} \sum_{i \in C_t} \mathbf{p}^t(i | \tilde{\mathcal{X}}_t) \hat{L}_{t,i}^2 \quad (47)$$

Proof: For the simplicity of the proof, fix t . Write $C = \tilde{\mathcal{X}}_t$. Let $p_i = \mathbf{p}_i^t$, $Z = Z_t$, $q_i = \mathbf{p}^t(i | \tilde{\mathcal{X}}_t)$, and $\hat{L}_i = \hat{L}_{t,i}$. Write the updated weights as $p_i^+ = p_i \exp(-\eta \hat{L}_i)$ and the updated active sum as $Z^+ = \sum_{i \in C} p_i^+$. Let $q_i^+ = p_i^+ / Z^+$. We denote the vector form of the aforementioned quantities in bold. e.g. $\mathbf{p}(i) = p_i$.

Note

$$\log \frac{q_i}{q_i^+} = \log \frac{p_i / Z}{p_i \exp(-\eta \hat{L}_i) / Z^+} = \log \frac{Z^+}{Z} + \eta \hat{L}_i \quad (48)$$

Now,

$$\begin{aligned} \text{KL}(\mathbf{v} \| \mathbf{q}^+) - \text{KL}(\mathbf{v} \| \mathbf{q}) &= \sum_{i \in C} v_i \log \frac{v_i}{q_i^+} - \sum_{i \in C} v_i \log \frac{v_i}{q_i} \\ &= \sum_{i \in C} v_i \log \frac{q_i}{q_i^+} \\ &= \log \frac{Z^+}{Z} + \eta \left\langle \mathbf{v}, \hat{\mathbf{L}} \right\rangle \end{aligned} \quad (49)$$

Rearranging we have

$$\eta \left\langle \mathbf{v}, \hat{\mathbf{L}} \right\rangle = \text{KL}(\mathbf{v} \| \mathbf{q}^+) - \text{KL}(\mathbf{v} \| \mathbf{q}) - \log \frac{Z^+}{Z} \quad (50)$$

Next, apply the fact that $e^{-x} \leq 1 - x + x^2$ for $x \geq 0$ and taking logs, we have

$$\begin{aligned} \log \frac{Z^+}{Z} &\leq \log \left(\sum_{i \in C} q_i \left(1 - \eta \widehat{L}_i + \frac{\eta^2}{2} \widehat{L}_i^2 \right) \right) \\ &= \log \left(1 - \eta \langle \mathbf{q}, \widehat{\mathbf{L}} \rangle + \frac{\eta^2}{2} \sum_{i \in C} q_i \widehat{L}_i^2 \right) \\ &\leq -\eta \langle \mathbf{q}, \widehat{\mathbf{L}} \rangle + \frac{\eta^2}{2} \sum_{i \in C} q_i \widehat{L}_i^2 \end{aligned} \quad (51)$$

where the last inequality applies $\log(1+x) \leq x$ for all $x > -1$.

Now plug (51) into (50),

$$\eta \langle \mathbf{v}, \widehat{\mathbf{L}} \rangle \geq \text{KL}(\mathbf{v} \parallel \mathbf{q}^+) - \text{KL}(\mathbf{v} \parallel \mathbf{q}) + \eta \langle \mathbf{q}, \widehat{\mathbf{y}} \rangle + \frac{\eta^2}{2} \sum_{i \in C} q_i \widehat{L}_i^2 \quad (52)$$

Rearranging the inequality and dividing by η , we get

$$\langle \mathbf{q} - \mathbf{v}, \widehat{\mathbf{L}} \rangle \leq \frac{1}{\eta} (\text{KL}(\mathbf{v} \parallel \mathbf{q}) - \text{KL}(\mathbf{v} \parallel \mathbf{q}^+)) + \frac{\eta}{2} \sum_{i \in C} q_i \widehat{L}_i^2 \quad (53)$$

Substituting the original notation back, we recover the claim. \square

G.4 REDUCTION TO BLOCKWISE FIXED ARM

The next two lemmas connect the one-step bound to a blockwise analysis under non-stationarity. Lemma 5 controls the discrepancy between the per-round best available arm and a single arm fixed over a block, showing that this gap is governed by the cumulative variation within the block. Lemma 6 then combines this control with the one-step OMD bound to obtain a regret bound against a fixed comparator over the block. Together, these results allow the per-round inequalities to be aggregated while isolating the effect of non-stationarity.

Lemma 5 Fix a block $\mathcal{B} = \{t_0, t_0 + 1, \dots, t_0 + L - 1\}$. Denote the blockwise best arm as

$$m_{\mathcal{B}} \in \arg \min_{i \in [K]} \widehat{L}_{t_0, i} \mathbb{I}\{i \in \tilde{\mathcal{X}}\} \quad (54)$$

Let Δ_t be as defined in Theorem 5. Then,

$$\sum_{t \in \mathcal{B}} \left(\widehat{L}_{t, m_{\mathcal{B}}} \mathbb{I}\{m_{\mathcal{B}} \in C_t\} - \widehat{L}_{t, m_t} \right) \leq L \sum_{t \in \mathcal{B} \setminus \{t_0\}} \Delta_t \quad (55)$$

Proof: Let $b_t(C) = \min_{i \in C} \widehat{L}_{t, i}$. By the Lipschitz property of min, it satisfies

$$|b_t(C) - b_{t-1}(C)| \leq \max_{i \in C} \left| \widehat{L}_{t, i} - \widehat{L}_{t-1, i} \right| \leq \Delta_t \quad (56)$$

By telescoping (56) from t_0 to t ,

$$b_{t_0}(C_t) \leq b_t(C_t) + \sum_{\tau=t_0+1}^t \Delta_\tau = m_t + \sum_{\tau=t_0+1}^t \Delta_\tau \quad (57)$$

By definition of $m_{\mathcal{B}}$ as the best static arm for the block

$$\sum_{t \in \mathcal{B}} \widehat{L}_{t, m_{\mathcal{B}}} \mathbb{I}\{m_{\mathcal{B}} \in \tilde{\mathcal{X}}\} \leq \sum_{t \in \mathcal{B}} \widehat{L}_{t, m_t^{t_0}} \mathbb{I}\{m_t^{t_0} \in \tilde{\mathcal{X}}\} \quad (58)$$

Also note that by definition,

$$\widehat{L}_{t, m_t^{t_0}} \leq \widehat{L}_{t_0, m_t^{t_0}} + \sum_{\tau=t_0+1}^t \Delta_\tau = b_{t_0}(\tilde{\mathcal{X}}) + \sum_{\tau=t_0+1}^t \Delta_\tau \quad (59)$$

1350 Combine (56) and (59),

$$1351 \sum_{t \in \mathcal{B}} \widehat{L}_{t, m_{\mathcal{B}}} \mathbb{I}\{m_{\mathcal{B}} \in \tilde{\mathcal{X}}_t\} \leq \sum_{t \in \mathcal{B}} \left(\widehat{L}_{t, m_t} + \sum_{\tau=t_0+1}^t \Delta_{\tau} \right) \quad (60)$$

1352 Rearranging, we get

$$1353 \sum_{t \in \mathcal{B}} \widehat{L}_{t, m_{\mathcal{B}}} \mathbb{I}\{m_{\mathcal{B}} \in \tilde{\mathcal{X}}_t\} - \widehat{L}_{t, m_t} \leq L \sum_{t \in \mathcal{B} \setminus \{t_0\}} \Delta_t \quad (61)$$

1354 \square

1355 **Lemma 6** Fix a block $\mathcal{B} = \{t_0, t_0 + 1, \dots, t_0 + L - 1\}$. Run Algorithm 2 with $\eta > 0$ starting from
1356 time t_0 . Let $\mathbf{q}_t \in \Delta_{\tilde{\mathcal{X}}_t}$ be the played distribution at each round $t \in \mathcal{B}$ and let $\widehat{\mathbf{L}}_t$ be an unbiased
1357 estimator of the masked loss vector

$$1358 \mathbb{E}[\widehat{\mathbf{L}}_t \mid \tilde{\mathcal{X}}_t] = \mathbf{L}_t^{\tilde{\mathcal{X}}_t}$$

1359 and that the second moment is bounded by

$$1360 \mathbb{E} \left[\sum_{i \in \tilde{\mathcal{X}}_t} \mathbf{q}_{t,i} \widehat{L}_{t,i}^2 \right] \leq \frac{k}{s} \quad (62)$$

1361 for all t . Then,

$$1362 \mathbb{E} \left[\sum_{t \in \mathcal{B}} \langle \mathbf{q}_t, \mathbf{l}_t^{C_t} \rangle - \mathbf{l}_{t, m_{\mathcal{B}}} \mathbb{I}\{m_{\mathcal{B}} \in \tilde{\mathcal{X}}_t\} \right] \leq \frac{\log(1/\alpha)}{\eta} + \frac{sk\eta L}{2} \quad (63)$$

1363 **Proof:** For each $t \in \mathcal{B}$ and any comparators $\mathbf{v}_t \in \Delta_{\tilde{\mathcal{X}}_t}$, apply Lemma 4

$$1364 \sum_{t \in \mathcal{B}} \langle \mathbf{q}_t - \mathbf{v}_t, \mathbf{L}_t \rangle \leq \frac{1}{\eta} \sum_{t \in \mathcal{B}} (\text{KL}(\mathbf{v}_t \parallel \mathbf{q}_t) - \text{KL}(\mathbf{v}_t \parallel \mathbf{q}_{t+1})) + \frac{\eta s^2}{2} \sum_{t \in \mathcal{B}} \sum_{i \in \tilde{\mathcal{X}}_t} \mathbf{q}_{t,i} \widehat{L}_{t,i}^2 \quad (64)$$

1365 By the unbiasedness of $\widehat{\mathbf{y}}_t$,

$$1366 \mathbb{E}[\langle \mathbf{q}_t - \mathbf{v}_t, \mathbf{l}_t \rangle] = \mathbb{E} \left[\langle \mathbf{q}_t - \mathbf{v}_t, \mathbf{l}_t^{\tilde{\mathcal{X}}_t} \rangle \right]$$

1367 For the KL terms, we have

$$1368 \sum_{t \in \mathcal{B}} (\text{KL}(\mathbf{v}_t \parallel \mathbf{q}_t) - \text{KL}(\mathbf{v}_t \parallel \mathbf{q}_{t+1})) = \text{KL}(\mathbf{v}_t \parallel \mathbf{q}_{t_0}) - \text{KL}(\mathbf{v}_t \parallel \mathbf{q}_{t_0+L-1}) \quad (65)$$

$$1369 \leq \log(1/\alpha)$$

1370 where the first equality is a result of telescoping with respect to t and the second due to the fact that
1371 \mathbf{q}_{t_0} is restarted from uniform.

1372 For the second term on the RHS of (64), we invoke Lemma 3. Taking expectation on both sides, we
1373 get (63). \square

1374 G.5 PROOF OF THEOREM 5

1375 Note that

$$1376 \sum_{t \in \mathcal{B}} \langle \mathbf{q}_t, \mathbf{l}_t^{\tilde{\mathcal{X}}_t} \rangle - \mathbf{l}_{t, m_t} = \left(\sum_{t \in \mathcal{B}} \langle \mathbf{q}_t, \mathbf{l}_t^{\tilde{\mathcal{X}}_t} \rangle - \mathbf{l}_{t, m_{\mathcal{B}}} \mathbb{I}\{m_{\mathcal{B}} \in \tilde{\mathcal{X}}_t\} \right) + \left(\sum_{t \in \mathcal{B}} \mathbf{l}_{t, m_{\mathcal{B}}} \mathbb{I}\{m_{\mathcal{B}} \in \tilde{\mathcal{X}}_t\} - \mathbf{l}_{t, m_t} \right) \quad (66)$$

1377 We bound the first term using Lemma 6 and the second term using 5. Taking the expectation and
1378 summing over blocks, we get

$$1379 \text{Reg}_n^{\text{BA}} = \mathbb{E} \left[\sum_{i=1}^B \sum_{t \in \mathcal{B}_i} \langle \mathbf{q}_t, \mathbf{l}_t^{\tilde{\mathcal{X}}_t} \rangle - \mathbf{l}_{t, m_t} \right] \leq \frac{T \log(1/\alpha)}{L} \frac{1}{\eta} + \frac{\eta s k}{2} T + L V_n \quad (67)$$

1404 Take

$$L^* = \left(\frac{T}{V_n} \sqrt{\frac{sk \log(1/\alpha)}{2}} \right)^{2/3}, \quad (68)$$

$$\eta^* = 2^{2/3} (\log(1/\alpha))^{1/3} (sk)^{-2/3} \left(\frac{V_n}{T} \right)^{1/3}. \quad (69)$$

1411 we get

$$\text{Reg}_n^{\text{BA}} \leq \frac{3}{2^{1/3}} (T^2 sk \log(1/\alpha) V_n)^{1/3}. \quad (70)$$

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458 H PERFORMANCE IMPROVEMENT CONTRIBUTION OF EACH PROBLEM

1459 We provide further justification for the per-problem policy improvement contribution defined in
 1460 Eq. (6). Intuitively, this quantity measures the marginal contribution of including a training problem \mathbf{x}
 1461 in an actor update to the overall improvement in policy performance under the evaluation distribution
 1462 $p_{\mathcal{X}}$.

1463 This interpretation is exact for a broad class of *tabular policies*, where the policy parameters for
 1464 different problems (or states) are independent. In such settings, updating the policy using trajectories
 1465 from a problem \mathbf{x} affects only the conditional distribution $\pi(\cdot | \mathbf{x})$ and leaves the policy unchanged
 1466 on all other problems.

1467 **Example: tabular REINFORCE.** Consider tabular REINFORCE, where for each problem \mathbf{x} the
 1468 policy $\pi(\cdot | \mathbf{x})$ is parameterized independently. At iteration t , suppose we update the policy using
 1469 rollouts collected *only* from a single problem \mathbf{x} . By construction, this update modifies $\pi(\cdot | \mathbf{x})$ but
 1470 does not change $\pi(\cdot | \mathbf{x}')$ for any $\mathbf{x}' \neq \mathbf{x}$.

1471 The overall performance objective is

$$1472 J(\pi) = \sum_{\mathbf{x}' \in \mathcal{X}} p_{\mathcal{X}}(\mathbf{x}') \mathbb{E}_{\mathbf{y} \sim \pi(\cdot | \mathbf{x}')} [R(\mathbf{y} | \mathbf{x}')].$$

1473 Since only the conditional policy at \mathbf{x} is changed, the performance difference $J(\pi^{t+1}) - J(\pi^t)$
 1474 depends solely on how the expected reward at \mathbf{x} changes.

1475 Applying the standard performance difference identity (Kakade & Langford, 2002) to the single-turn
 1476 setting yields

$$1477 J(\pi^{t+1}) - J(\pi^t) = p_{\mathcal{X}}(\mathbf{x}) \mathbb{E}_{\mathbf{y} \sim \pi^t(\cdot | \mathbf{x})} \left[\frac{\pi^{t+1}(\mathbf{y} | \mathbf{x})}{\pi^t(\mathbf{y} | \mathbf{x})} A_{\pi^t}(\mathbf{y} | \mathbf{x}) \right],$$

1478 which is exactly the per-problem utility $u_{\mathbf{x}}^t$ defined in Eq. (6). Thus, in tabular REINFORCE, updating
 1479 the policy on a single problem \mathbf{x} produces an expected performance improvement of $u_{\mathbf{x}}^t$.

1480 The same reasoning extends directly to batch updates: if the policy is updated using a subset of
 1481 problems \mathcal{X}^t , the total performance improvement decomposes additively as $\sum_{\mathbf{x} \in \mathcal{X}^t} u_{\mathbf{x}}^t$, and including
 1482 a problem \mathbf{x} in the update contributes exactly $u_{\mathbf{x}}^t$ to the expected performance gain.

1483 **Other tabular methods.** This exact additive interpretation applies equally to other tabular rein-
 1484 forcement learning methods, including tabular policy gradient methods, tabular Q-learning, and
 1485 SARSA, where updates based on a problem or state affect only the corresponding local policy or
 1486 value parameters. In all such cases, the per-problem utility $u_{\mathbf{x}}^t$ captures the true marginal contribution
 1487 of training on \mathbf{x} .

1488 **Discussion: function approximation.** In the presence of function approximation, updating the
 1489 policy using trajectories from one problem generally affects the policy on other problems as well,
 1490 breaking the exact additivity described above. Analyzing such cross-problem interference requires
 1491 strong assumptions on the structure of the function class and the optimization dynamics, and is
 1492 beyond the scope of this work.

1493 Nevertheless, when policy updates are small—as is typical in modern RL post-training algorithms
 1494 such as PPO-style methods, GRPO, or GSPO—the per-problem utility $u_{\mathbf{x}}^t$ remains a first-order
 1495 approximation to the marginal contribution of problem \mathbf{x} to performance improvement. Our empirical
 1496 results in Sec. 4.1 indicate that this approximation is sufficiently accurate to drive effective curriculum
 1497 learning at scale.

1512 I LIMITATIONS
1513

1514 While ACTOR-CURATOR is agnostic to the specific actor update rule and can be combined with
1515 a wide range of post-training algorithms, its gains ultimately depend on the stability and quality
1516 of the underlying actor optimization: unstable updates can still lead to noisy dynamics or training
1517 collapse, a common failure mode in reinforcement learning that curriculum learning alone cannot
1518 fully address. In addition, ACTOR-CURATOR inherits standard assumptions from RL-based LLM
1519 post-training, most notably access to a reliable reward signal, making it best suited to domains with
1520 objectively verifiable rewards such as mathematics, logic puzzles, and code. Finally, learning an
1521 explicit curator introduces additional computational overhead beyond standard RL post-training;
1522 while modest relative to actor training and rollout generation, this cost is non-negligible, and in our
1523 experiments ACTOR-CURATOR increases overall wall-clock training time by approximately 9% (see
1524 App. K), though this overhead is small compared to the observed gains in training efficiency.

1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

J EXPERIMENTAL SETUP

J.1 HYPER-PARAMETERS

Unless otherwise specified, all experiments use the same hyper-parameter configuration for ACTOR-CURATOR. The complete set of hyper-parameters is summarized in Table 2. We briefly explain key parameters below, with a focus on those specific to the actor–curator framework.

Parameter	Value
<i>Model configuration</i>	
Curator model	Qwen3-0.6B
Actor KL loss	Disabled
Max problem length	1024 tokens
Max solution length	4096 tokens
<i>Sampling and batch sizes</i>	
Candidate batch size $ \tilde{\mathcal{X}}^t $	2048
Training batch size $ \mathcal{X}^t $	256
Rollouts per problem $ \mathcal{Y}_x^t $	8
<i>Actor optimization</i>	
Actor temperature	1.0
Actor training top- p	1.0
Actor validation top- p	0.7
Actor learning rate	1×10^{-6}
Actor LR warmup ratio	0.05
Actor weight decay	0.1
Actor gradient clipping	1.0
Actor clip range ρ_{\min}, ρ_{\max}	$[3 \times 10^{-4}, 4 \times 10^{-4}]$
<i>Curator optimization</i>	
Curator dormant steps	20
Curator warm-up steps	5
Curator temperature	1.0
Curator top- p	0.9
Curator learning rate	1×10^{-6}
Curator PPO clip range $\rho_{\max} - \rho_{\min}$	0.2

Table 2: Hyper-parameters used for ACTOR-CURATOR across all experiments unless otherwise specified.

Candidate and training batch sizes. At each training iteration, a candidate batch $\tilde{\mathcal{X}}^t$ of size 2048 is first sampled from the proposal distribution q . The curator then reweights this candidate set and samples a smaller training batch \mathcal{X}^t of size 256, which is used for actor rollouts and updates. This two-stage sampling scheme follows Section 3.5 and allows scalable curation over large problem banks.

Rollouts per problem. For each selected problem $x \in \mathcal{X}^t$, we generate $|\mathcal{Y}_x^t| = 8$ on-policy rollouts from the current actor. These rollouts are used both for the actor update and for estimating per-problem policy improvement signals used to train the curator.

Curator dormant and warm-up steps. During the first *curator dormant steps* (20 iterations), problems are sampled uniformly from the candidate batch, and curator outputs are ignored. This stabilizes early actor learning before meaningful policy-improvement estimates can be obtained. During the subsequent *curator warm-up steps* (5 iterations), the curator begins to influence sampling, but its parameters are updated conservatively. After warm-up, the curator is fully active and trained online using bandit feedback.

Sampling prior. When use `sampling prior` is enabled, curator-assigned weights are multiplied by the proposal-induced marginal inclusion probability $q(x)$ before normalization. This

1620 encourages coverage of the full dataset and prevents the curator from collapsing onto a narrow subset
1621 of problems early in training, consistent with the two-stage unbiased estimator in Equation (12).
1622

1623 **Proximal curator clipping.** Curator updates use the PPO-style clipped OSMD objective described
1624 in Section 3.6. The clipping range ρ_{\min}, ρ_{\max} constrains the importance ratio between consecutive cu-
1625 rator policies, stabilizing learning under function approximation. The additional clip range parameter
1626 controls the maximum allowed deviation between these bounds.
1627

1628 J.2 HARDWARE

1629
1630 All experiments were conducted on NVIDIA A100 and H200 GPUs. Actor rollout generation and
1631 optimization dominate overall runtime; curator training introduces approximately 14% additional
1632 wall-clock cost relative to uniform sampling, as discussed in Appendix E.
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673

1674 K ADDITIONAL ANALYSIS

1675

1676 K.1 OVERHEAD

1677

1678 We found that ACTOR-CURATOR adds about 9% training wall time overhead. We present the
1679 overhead by dataset and model in Tab. 3. However, as shown in Fig. 11, this is relatively the compared
1680 to the efficiency gains.

1681

1682 Table 3: **Average wall-time overhead percentage** by model and datasets, averaged over training
1683 steps 500 steps.

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

	COUNTDOWN	ZEBRA	ARC-1D	MATH
QWEN2.5-3B-BASE	11.86	16.71	17.12	9.56
LLAMA3.2-3B-IT	9.34	13.63	19.40	9.82

L ADDITIONAL ABLATION

Candidate batch size $|\tilde{\mathcal{X}}|$. As shown in Fig. 8, candidate sizes of 512 and 2048 yield similar final performance. Larger batches (e.g., 8192) lead to unstable training, likely due to reduced exploration and overfitting.

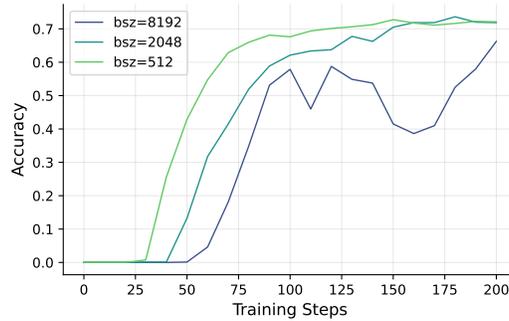
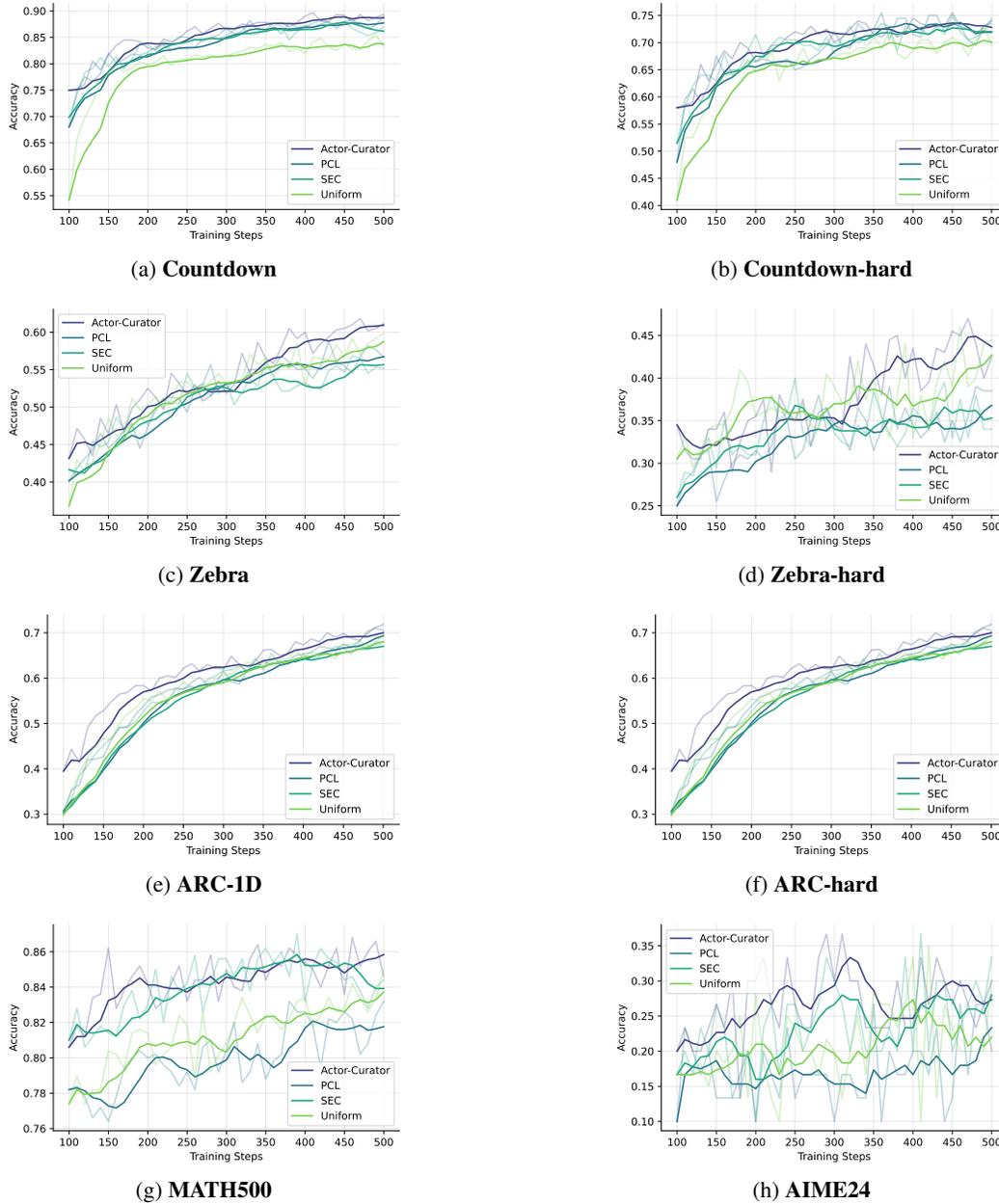


Figure 8: Effect of candidate batch size on test performance. While `bsz=2048` and `bsz=512` converge to similar performance at step 200, `bsz=8192` suffers from instability. Sampled batch size is 256.

1782 Figure 9: Test set performance across training steps (within 100 training steps) for **Qwen2.5-3B-Base**
 1783 across benchmarks and methods.
 1784

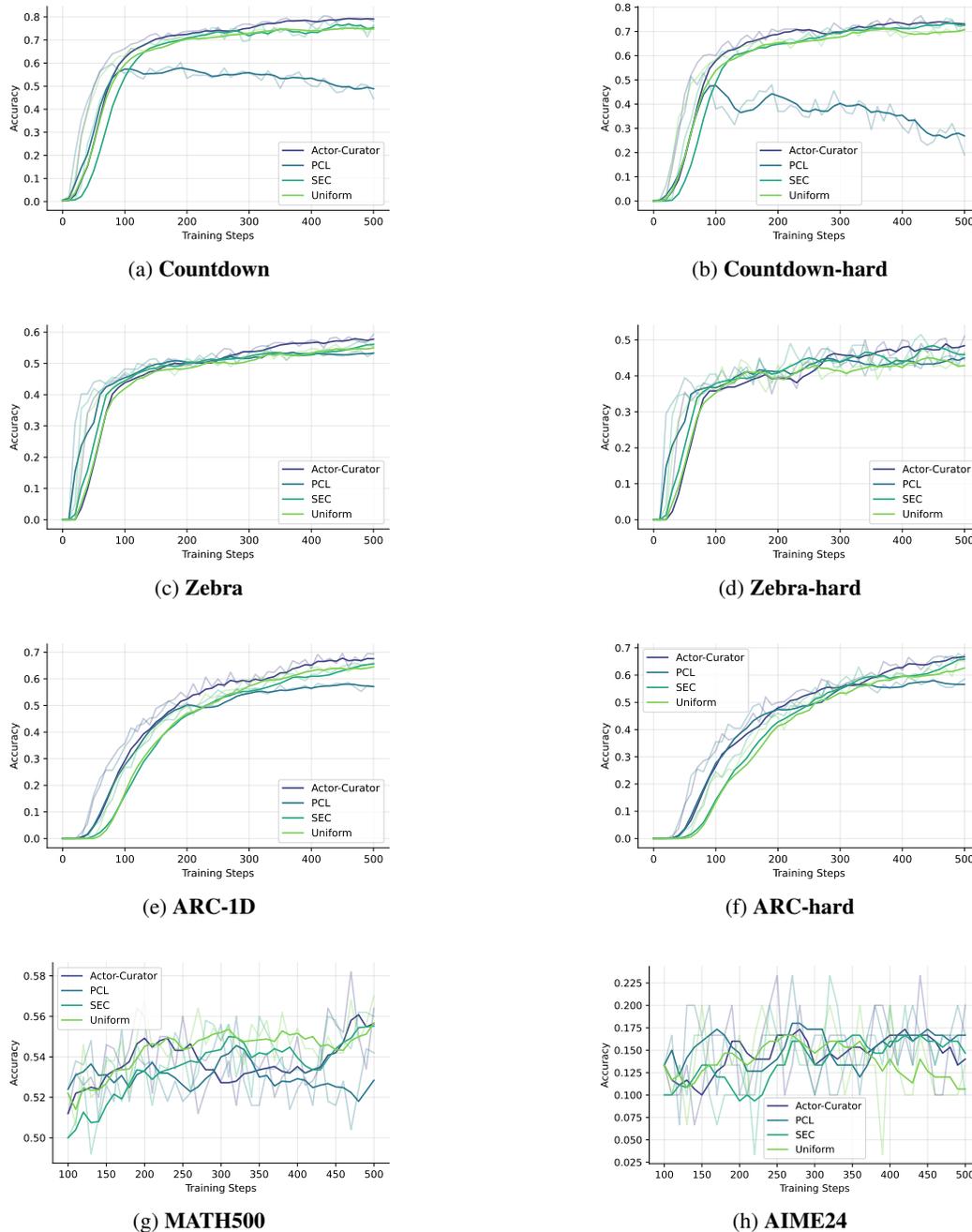


1827 M MAIN RESULTS (EXTENDED)

1829 This appendix provides extended empirical results complementing the main paper. We report
 1830 additional quantitative comparisons across benchmarks, models, and curriculum learning methods,
 1831 as well as detailed training dynamics over time. These results further substantiate the robustness,
 1832 efficiency, and stability of AC across diverse problem domains and model backbones.
 1833

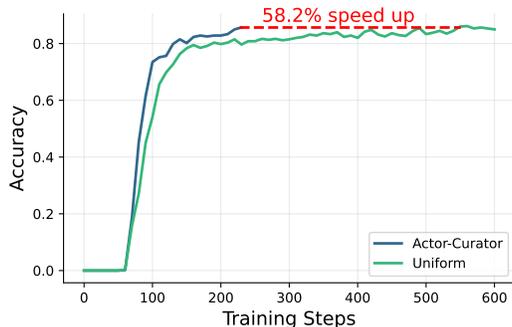
1834 **Extended performance comparison.** Tab. 4 reports peak validation performance within the first
 1835 100 training steps across all benchmarks and models. Compared to uniform sampling, heuristic curricula (SEC), and learning-based baselines (PCL), AC consistently achieves higher peak performance

1836 Figure 10: Test set performance across training steps (within 100 training steps) for **Llama3.2-3B-it**
 1837 across benchmarks and methods.
 1838

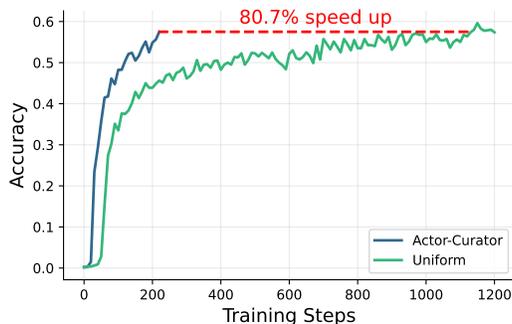


1886 on most benchmarks. The gains are particularly pronounced on harder subsets (e.g., Countdown-
 1887 hard, Zebra-hard, ARC-hard, and AIME24), highlighting the effectiveness of directly optimizing for
 1888 expected policy improvement when problem difficulty and utility are highly non-uniform. While per-
 1889 formance on MATH500 is largely saturated for some model configurations, AC remains competitive
 and avoids degradation relative to strong baselines.

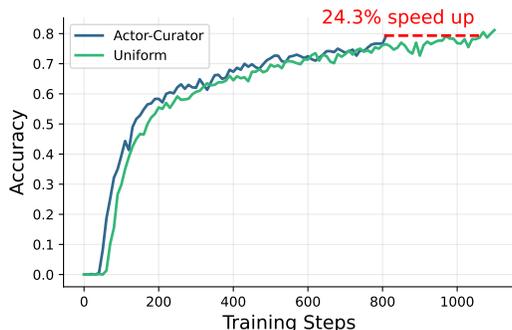
1890 **Figure 11: Training efficiency:** Actor-Curator attains significant efficiency increase with relatively
 1891 low overhead on Countdown, Zebra, and ARC.



(a) Countdown



(b) Zebra



(c) ARC-1D

1934 **Training dynamics and stability.** Figs. 9 and 10 visualize test performance as a function of training
 1935 steps for Qwen2.5-3B-Base and Llama3.2-3B-it, respectively. Across benchmarks, AC not only
 1936 reaches higher peak performance but also exhibits faster convergence and more stable learning
 1937 dynamics. In many cases, competing methods plateau early or exhibit higher variance, whereas AC
 1938 continues to make steady progress, effectively raising the performance ceiling.

1939
1940 **Model-agnostic behavior.** The trends observed in Figs. 9 and 10 are consistent across both base
 1941 and instruction-tuned models, indicating that the benefits of AC are not tied to a specific initialization
 1942 or training regime. This supports the claim that learning curricula via policy-improvement-driven
 1943 signals provides a generally applicable mechanism for improving RL post-training efficiency and
 robustness.

1944 Table 4: Peak validation performance on problems within 100 training steps across models and
 1945 methods. ACTOR-CURATOR outperforms both other learning based methods (PCL) and methods that
 1946 rely on human heuristics (SEC)
 1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

BENCHMARK	MODEL	METHOD					IMPROVEMENT	
		π_{ref}	UNIFORM	SEC	PCL	AC (OURS)	+ Δ	+ $\Delta\%$
COUNTDOWN	QWEN2.5-3B-BASE	0.00	44.74	58.87	57.24	62.12	+3.25	+5.52
	LLAMA3.2-3B-IT	0.00	63.12	62.78	59.62	66.25	+3.13	+4.96
COUNTDOWN-HARD	QWEN2.5-3B-BASE	0.00	41.00	51.50	48.00	58.00	+6.50	+12.62
	LLAMA3.2-3B-IT	0.00	58.50	58.00	53.00	60.50	+2.00	+3.42
ZEBRA	QWEN2.5-3B-BASE	0.00	35.12	36.00	34.12	37.62	+1.62	+4.50
	LLAMA3.2-3B-IT	0.00	44.50	46.50	48.25	47.12	-1.13	-2.34
ZEBRA-HARD	QWEN2.5-3B-BASE	0.00	30.50	27.50	26.00	34.50	+4.00	+13.11
	LLAMA3.2-3B-IT	0.00	37.50	38.00	39.50	40.50	+1.00	+2.53
ARC-1D	QWEN2.5-3B-BASE	0.00	26.74	27.87	26.37	36.37	+8.50	+30.51
	LLAMA3.2-3B-IT	0.00	27.62	26.75	34.50	35.25	+0.75	+2.17
ARC-HARD	QWEN2.5-3B-BASE	0.00	19.50	18.50	18.50	31.00	+11.50	+58.97
	LLAMA3.2-3B-IT	0.00	23.00	24.50	24.00	31.50	+7.00	+28.57
MATH500	QWEN2.5-3B-BASE	61.80	83.00	81.00	79.79	81.00	-2.00	-2.41
	LLAMA3.2-3B-IT	41.00	52.20	52.00	52.40	53.60	+1.20	+2.29
AIME24	QWEN2.5-3B-BASE	3.33	23.33	20.00	23.33	30.00	+6.67	+28.57
	LLAMA3.2-3B-IT	0.00	13.33	13.33	13.33	16.67	+3.34	+25.06

1968 **Training efficiency.** Fig. 11 compares learning curves of Actor-Curator against uniform sampling
 1969 on Countdown, Zebra, and ARC. Across all three benchmarks, Actor-Curator reaches the same
 1970 target accuracy substantially earlier, yielding step-level speedups of 58.2% on Countdown, 80.7% on
 1971 Zebra, and 24.3% on ARC. This indicates that policy-improvement-driven data selection primarily
 1972 accelerates optimization by prioritizing high-impact problems.
 1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997