
Frontier-Weighted SEC: A Case Study in Curriculum Learning for RL Fine-Tuning of Language Models

Garrett Alarcon*
Stanford University
gaa@stanford.edu

Extended Abstract

Motivation & problem. Online policy-gradient methods such as RLOO Ahmadian et al. [2024] are increasingly standard for post-training language models on verifiable reasoning tasks DeepSeek-AI [2025], Pan et al. [2025], but RL fine-tuning of LLMs is compute-expensive. Curriculum learning Bengio et al. [2009] is an appealing lever: if the policy can be trained on the “right” difficulty mix, fewer GPU-hours might reach the same end performance. The Self-Evolving Curriculum (SEC) of Chen et al. [2025] casts curriculum as a non-stationary multi-armed bandit over difficulty bins, with per-bin reward equal to mean absolute advantage. SEC was demonstrated on Qwen-3B/7B with 120–240 RLOO steps; its behavior in the smaller-model and shorter-budget regime typical of academic projects has not been characterized.

Method. We replicate SEC on Countdown 3-to-4 with Qwen2.5-0.5B and 100 RLOO steps, report an empirical regularity in its reward signal, and propose one algorithmic modification. The contributions are: (i) during RLOO training under our settings, mean $|\text{adv}|$ on 3-operand bins consistently exceeds 4-operand bins by 0.07 ± 0.02 (mean \pm std across one v3 run and three v4 seeds with very different bandit sampling distributions), opposite the direction curriculum learning intends since 4-operand bins are 2–4 \times harder by test reward; (ii) a *frontier-weighted* reward signal $r_{\text{bin}} = |\text{adv}| + \lambda \cdot \max(1 - \bar{r}, 0) \cdot \mathbf{1}[\bar{r} > 0]$ shifts the bandit’s per-bin sampling toward harder bins (53/47 easy/hard sampling under standard SEC becomes 35/65 under our variant, as logged from per-bin visit counts); (iii) at our scale the redirected curriculum does not outperform uniform sampling on test reward. We do not propose a confirmed mechanism for the $|\text{adv}|$ asymmetry; our post-hoc analysis of eval-time reward distributions does not predict it.

Results. We implemented the assigned components of SFT, IPO, and RLOO (the `NotImplementedError` blocks in the project-provided starter code) and reproduced the teaching baselines: SFT 0.374, IPO 0.402, RLOO uniform baseline 0.535, each above its CS224R autograder threshold. With three random seeds at $\lambda = 1$, our frontier-weighted SEC achieves 0.513 ± 0.015 versus the baseline’s 0.535 (single seed). A one-sample t -test against the baseline gives $t = -2.56$ on $df = 2$, which does not reject the null of equality at $p = 0.05$; we cannot demonstrate improvement or equivalence at this seed budget.

Discussion, limitations, conclusion. Principal limitations: only 3 seeds for v4 and 1 for the baseline (no paired test); single model size and budget; the training-time $|\text{adv}|$ asymmetry is not isolated by our data (the eval-time reward distribution does not predict it, suggesting a training-time-specific interaction of sampling temperature, policy state, and reward structure). Future work: a refined learnability gate (our $\mathbf{1}[\bar{r} > 0]$) is effectively always active in our setup because Countdown is highly format-compliant), a strict binary-reward ablation, and replication at larger model sizes and longer budgets. Practitioners considering SEC for sparse-reward LLM tasks should measure per-bin $|\text{adv}|$ early in training and confirm it scales with the intended curriculum direction before adopting the standard SEC signal; the diagnostic is cheap.

*Solo team. gaa@stanford.edu, Stanford CGOE.

Abstract

Curriculum learning via multi-armed bandits is increasingly used to allocate RL training compute across problems of varying difficulty. The Self-Evolving Curriculum (SEC) of Chen et al. (2025), which uses mean absolute advantage as the per-bin reward signal, has been demonstrated on Qwen-3B/7B with hundreds of RLOO steps. We investigate SEC in the smaller-model, shorter-budget regime typical of course projects: Qwen2.5-0.5B and 100 RLOO steps on Countdown 3-to-4. During training, the mean $|\text{adv}|$ signal is larger on easier 3-operand bins than on harder 4-operand bins despite the harder bins being where the policy has more room to improve. The asymmetry persists across four training runs with very different sampling distributions ($\sim +0.07$ in favor of 3-operand bins). We propose a *frontier-weighted* reward signal that adds a “room to grow” bonus; per-bin bandit logs show the bandit shifts toward harder bins under this signal. Across three seeds, the redirected curriculum does not statistically outperform uniform RLOO on test reward. We report this as a case study: an empirical regularity in SEC’s reward signal in a previously uncharacterized regime, an algorithmic modification whose effect on bandit sampling is measurable, and a negative result on overall task performance, with implications for the design of curriculum bandits on sparse-reward LLM fine-tuning tasks.

1 Introduction

RL fine-tuning of pretrained language models is now the dominant approach for eliciting reasoning behavior DeepSeek-AI [2025], Pan et al. [2025], but it is expensive: each rollout costs a forward pass through the policy, and verifier-based RL typically generates many rollouts per prompt to obtain a useful gradient signal. *Curriculum learning* — exposing the policy to a distribution of training problems calibrated to its current ability — is an appealing way to reduce this cost. If the curriculum concentrates training data on the right “learning frontier,” the same end-task performance might be reachable in fewer gradient steps, or higher performance reachable in the same budget.

Among curriculum learning approaches for RL fine-tuning, the Self-Evolving Curriculum (SEC) Chen et al. [2025] is attractive because it is fully automatic: it treats curriculum allocation as a non-stationary multi-armed bandit over difficulty bins, requires no separate model of student ability, and the per-bin reward signal (mean absolute advantage) is already available as a by-product of standard policy-gradient training. SEC has been shown to outperform uniform sampling on Countdown, ARC, and MATH at Qwen-3B and 7B scale with 120–240 RLOO steps.

It is not a priori clear whether SEC’s positive results transfer to the smaller-model, shorter-budget regimes that course projects must use. Smaller models exhibit different reward distributions (more failure modes, sparser successes) and shorter training budgets give curriculum less opportunity to compound. In this work we replicate SEC on top of RLOO for Countdown 3-to-4 with Qwen2.5-0.5B and 100 RLOO steps, and ask:

- **RQ1.** Does SEC’s mean $|\text{adv}|$ reward signal correctly identify the learning frontier in a small-model, sparse-reward setting?
- **RQ2.** If not, can the bandit be redirected with a modified reward signal, and is the redirection verifiable independent of test reward?
- **RQ3.** Does correctly directing the curriculum improve end-task performance at this scale?

We answer **RQ1** negatively (§3.3): training-time $|\text{adv}|$ is higher on easier bins than on harder ones in our regime, opposite the intended direction. The asymmetry persists across runs with very different bandit sampling distributions (oversampling easy bins in v3 and oversampling hard bins in all three v4 seeds). The precise mechanism is not isolated by our data. We answer **RQ2** positively (§3.4, §5): a frontier-weighted reward signal shifts per-bin sampling toward harder bins, as measured from per-bin visit counts. We answer **RQ3** negatively (§5): the redirected curriculum produces no statistically distinguishable improvement on test reward at our budget (one-sample $t = -2.56$, $df = 2$, fails to reject equality at $p = 0.05$), and our 300-step budget probe (§??) suggests both methods over-train past 100 steps with the standard KL setting.

This work is a case study: we reproduce the teaching baselines from the project’s starter code, identify an empirical regularity in SEC’s reward signal in a previously uncharacterized regime, and explore one algorithmic modification. We do not propose a new SOTA curriculum bandit, and we do not claim a mechanistic explanation for the regularity — only that the regularity exists, persists across the obvious sampling confounds, and is in principle correctable at the bandit level.

2 Related Work

Curriculum learning for RL. Automatic curriculum learning has a long history in deep RL, with methods that pick training tasks based on signals like absolute learning progress Portelas et al. [2019], value disagreement, or competence-based heuristics. For LLMs, the dominant recent approach is SEC Chen et al. [2025], which casts difficulty bins as arms of a non-stationary multi-armed bandit. SEC’s reward signal — mean absolute advantage per bin — is justified by the Zone of Proximal Development principle: under binary rewards, $|\text{adv}|$ is theoretically maximized at success rate $p = 0.5$, the same point at which curriculum learning theory predicts maximum information gain. Our case study differs from Chen et al. [2025] in three ways: a smaller model (0.5B vs 3B/7B), a shorter budget (100 vs 120–240 RLOO steps), and a focus on diagnosing failure modes rather than demonstrating performance gains.

Preference-based and verifier-based LLM fine-tuning. Preference-based methods like DPO Rafailov et al. [2023] and IPO Azar et al. [2024] reparameterize human preference learning as a supervised classification objective, avoiding the need for an explicit reward model. Verifier-based RL methods like RLOO Ahmadian et al. [2024] instead use online rollouts scored by a rule-based verifier, which is well-suited to math, code, and reasoning tasks. Our pipeline implements both: IPO as an offline alignment stage on a preference dataset, RLOO as an online policy-gradient stage on a verifier dataset. SEC sits on top of RLOO and modifies only the prompt-sampling distribution.

Reward shaping under sparse reward. The challenge of sparse, ternary, or otherwise structured rewards is widely discussed in the policy-gradient literature: advantage collapse, high-variance gradients, and reward hacking are common failure modes. The empirical observation we report — a directional $|\text{adv}|$ asymmetry across operand-count tiers in our Countdown / Qwen-0.5B setup that opposes the curriculum’s intent — may have been characterized in prior work we have not surveyed. We offer it as a case study observation rather than a definitive novelty claim, and we discuss its plausible determinants (sampling temperature, policy state, ternary reward structure) in §3.3.

3 Method

3.1 Background: RLOO + SEC

RLOO policy gradient. For a prompt x we sample G rollouts $y^{(1)}, \dots, y^{(G)}$ from the current policy π_θ and score each with a verifier $R(\cdot)$. The leave-one-out advantage is $A^{(i)} = \frac{G \cdot R(y^{(i)}, x) - \sum_j R(y^{(j)}, x)}{G-1}$. Because vLLM Kwon et al. [2023] (behavior policy μ) and Hugging-Face (target policy π_θ) compute token log-probabilities with slightly different kernels, we apply per-sample importance weighting $w = \exp(\log \pi_\theta(y|x) - \log \mu(y|x))$, computed in log-space and clipped to a max log-ratio of 5. The combined objective minimizes

$$\mathcal{L} = -\mathbb{E}[w \cdot A \cdot \log \pi_\theta(y|x)] - \alpha_{\text{ent}} \mathcal{H}[\pi_\theta] + \beta_{\text{KL}} D_{\text{KL}}^{K3}(\pi_\theta \parallel \pi_{\text{ref}}), \tag{1}$$

where D_{KL}^{K3} is the Schulman K3 estimator and π_{ref} is a frozen reference (the SFT checkpoint).

SEC bandit. SEC partitions training prompts into K difficulty bins. Each bin b has a Q-value $Q(b)$ updated by TD(0) EMA: $Q(b) \leftarrow (1 - \alpha)Q(b) + \alpha \cdot r_b$, where r_b is the per-round reward signal for bin b . The next batch’s bins are sampled from a Boltzmann distribution: $P(b) \propto \exp(Q(b)/T)$. The standard SEC reward signal is $r_b = \text{mean}_{i \in b} |A^{(i)}|$, i.e. mean absolute advantage of rollouts whose prompt was drawn from bin b . The Zone of Proximal Development argument is that, under binary rewards with group size G , this signal peaks at per-bin success rate $p = 0.5$.

3.2 Bug fix found en route: pad-token contamination in importance weight

While building the RLOO baseline, we observed importance weights collapsing to $w \approx 10^{-4}$ within ten training steps and rollout reward degrading from 0.33 to 0.18. The root cause was a mask error in the RLOO batch construction: after right-padding response token sequences to a common length, the original code marked *every column of the response block* as a response token via `np.ones_like(response_input_ids)` — including the pad columns added by the HuggingFace tokenizer. The HF log-probability $\log \pi_\theta$ was therefore summed over real response tokens *plus* pad tokens, while vLLM’s `cumulative_logprob` (used as $\log \mu$) covered only generated tokens. The unit mismatch drove $\log \pi_\theta - \log \mu \approx -8$ for typical responses. Replacing `np.ones_like(...)` with the actual `response_attention_mask` restored the correct unit. After the fix, the mean importance weight is in the right regime ($\sim 1\text{--}3$) and KL to the reference stays bounded (Figure 1). All numbers reported in this paper use the fixed RLOO implementation.

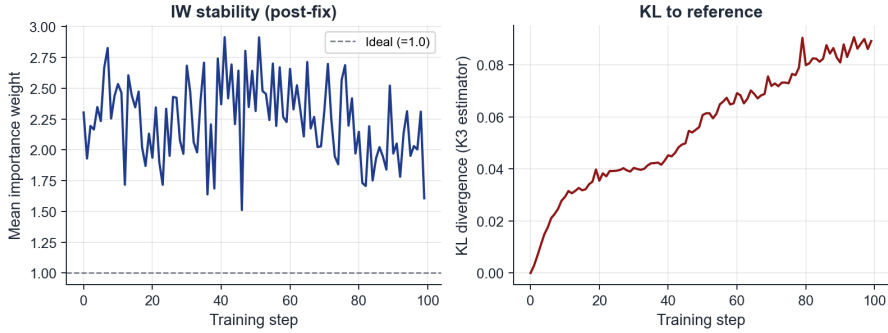


Figure 1: Importance weight (left) and K3 KL divergence (right) for the uniform RLOO baseline at 100 steps after fixing the pad-token mask. w stays in the $[1, 3]$ range; KL grows monotonically but does not explode.

3.3 Diagnosis: $|\text{adv}|$ asymmetry on Countdown

We binned the Countdown training set by composite key $(n_{\text{operands}}, q_{\text{target}})$ where q_{target} is the quantile bucket of the target value (4 buckets, edges at empirical quantiles of the train-set target distribution). This produces 8 bins. We then ran SEC v3 (8 composite bins, Boltzmann temperature $T = 0.3$, UCB exploration bonus, mean $|\text{adv}|$ reward) and measured per-bin $|\text{adv}|$ during training.

Empirical observation. The mean per-bin $|\text{adv}|$ averaged 0.246 across 3-operand bins and 0.172 across 4-operand bins under SEC v3 (Figure 2). This asymmetry is consistent across all four SEC runs in our dataset (one v3 and three v4 seeds), with a magnitude of 0.072 ± 0.016 in favor of the 3-operand tier (Figure 3). Notably, the asymmetry persists even under SEC v4, which redirects the bandit to oversample hard bins (so 4-operand bins receive *more* training data than 3-operand bins under v4) — ruling out sampling imbalance as a confound. Because 4-operand bins are 2–4 \times harder than 3-operand bins by baseline test reward (§5), the standard SEC bandit (which treats $|\text{adv}|$ as the per-arm reward) allocates more attention to bins the policy has already largely solved.

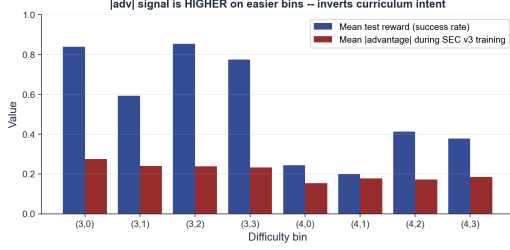


Figure 2: Per-bin mean reward (blue, baseline test set) and per-bin mean |adv| during SEC v3 training (red). 3-operand bins have higher training |adv| despite lower difficulty.

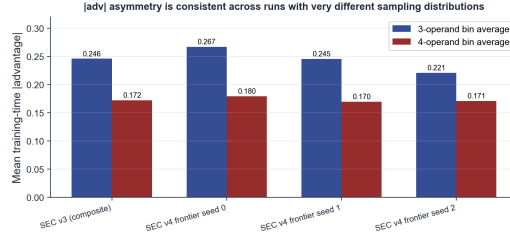


Figure 3: Mean training-time |adv| on 3-operand vs. 4-operand bins across four training runs with very different bandit sampling distributions. The asymmetry direction and magnitude is consistent.

Mechanism is not isolated. We initially conjectured that the asymmetry was caused by Countdown’s ternary verifier reward $\{0.0, 0.1, 1.0\}$: rollouts on hard bins would cluster at the 0.1 “format-only” partial credit, compressing within-group reward variance and suppressing |adv|. *Our post-hoc analysis does not strongly support this specific mechanism.* The eval-time per-bin reward distribution on the trained baseline (Figure 4) shows that hard 4-operand bins are dominated by 0.0 (no answer) rather than 0.1 (format-only). Computing expected mean |adv| directly from these distributions yields nearly identical values for 3-operand and 4-operand bins (~ 0.21 each); the training-time asymmetry we observe (~ 0.27 vs. ~ 0.17) does not appear at eval time (Figure 5). We therefore believe the asymmetry is a training-time phenomenon that depends on some combination of (a) sampling temperature (training uses $T = 1.0$ with no top- k /top- p filtering per the project spec, vs. $T = 0.6$ at eval), (b) policy state and the trajectory through which the policy improves on each bin during training, and (c) the ternary reward structure interacting with the above; we are unable to isolate any of these factors from our data. We treat the asymmetry as an empirical regularity and the v4 fix as targeting its symptom rather than a confirmed root cause.

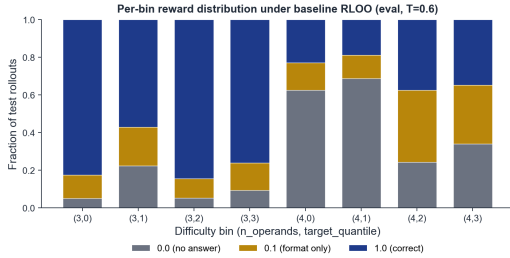


Figure 4: Per-bin reward distribution on the baseline-trained policy at eval. Hard 4-operand bins cluster at 0.0 (gray), not at 0.1 (gold) as we originally hypothesized.

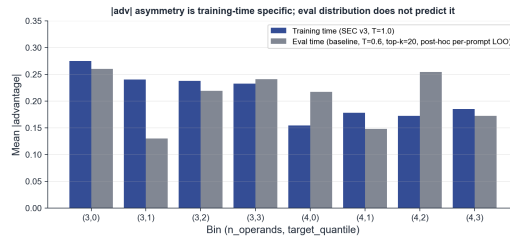


Figure 5: Training-time vs. eval-time per-bin mean |adv|. The asymmetry is sharp at training time but largely absent at eval, suggesting training-time-specific factors.

3.4 Proposed fix: frontier-weighted reward signal (SEC v4)

We modify the SEC per-bin reward signal to add an explicit “room to grow” bonus, gated on non-zero success rate:

$$r_{\text{bin}} = \underbrace{|\text{adv}|}_{\text{informative}} + \lambda \cdot \underbrace{\max(1 - \bar{r}, 0)}_{\text{room to grow}} \cdot \underbrace{\mathbf{1}[\bar{r} > 0]}_{\text{feasible}}, \quad (2)$$

where \bar{r} is the bin’s mean rollout reward in the current batch. The three factors encode three different curriculum desiderata. |adv| captures the standard SEC “training is informative here” signal. $\max(1 - \bar{r}, 0)$ adds the “policy has room to improve here” signal. The gate $\mathbf{1}[\bar{r} > 0]$ is intended to suppress bins where the policy never succeeds, preventing over-allocation to infeasible bins.

Practical note on the gate. In our Countdown setup, with batch size 128 prompts and $G = 8$ rollouts per prompt, each bin contributes $\approx 16 \times 8 = 128$ rollouts per training round under uniform sampling. Combined with the fact that the SFT-warm-started policy emits parseable answers on almost all prompts (format compliance $> 99\%$ in our baseline run), the bin-level mean reward $\bar{r} > 0$ in essentially every round on essentially every bin. Our gate is therefore effectively always active and the signal reduces to $r_{\text{bin}} = |\text{adv}| + \lambda(1 - \bar{r})$ in practice. We retain the gate term in the formula for completeness (it would matter on tasks where the policy genuinely cannot solve the hardest bins) and discuss in §6 how a non-trivial success-rate threshold would be more useful in our regime.

What is novel. We do not change SEC’s bandit structure (still Boltzmann over EMA-Q), its difficulty binning, or its update rule. The only change is the per-bin reward signal in Equation 2. The signal is a simple combination of two terms that are individually standard in the curriculum learning literature: informativeness via $|\text{adv}|$ from Chen et al. [2025] and a learning-progress proxy via $1 - \bar{r}$ analogous in spirit to Portelas et al. [2019]. Our contribution is the observation that the standard SEC signal misallocates on Countdown 3-to-4 at our scale, plus the specific reweighting and its (verifiable) effect on bandit sampling.

4 Experimental Setup

Task and datasets. We use the Countdown 3-to-4 arithmetic reasoning task from Pan et al. [2025]. Each prompt provides 3 or 4 integers and a target value; the policy must emit an arithmetic equation using each number at most once that evaluates to the target. Datasets: SFT warm-start on `Asap7772/cog_behav_all_strategies`; IPO on `asingh15/countdown_tasks_3to4-dpo`; RLOO on `asingh15/countdown_tasks_3to4` (verifier-based). The reward function (in `evaluation/countdown.py`) returns 0.0 for no `<answer>` tag, 0.1 for a parseable wrong answer, and 1.0 for a correct answer.

Model. Qwen2.5-0.5B-Base for all experiments, per the project’s stated constraint.

Pipeline baseline hyperparameters. SFT: 6 epochs, AdamW lr 5×10^{-5} , batch 64, gradient accumulation 8, gradient clipping 1.0, warmup ratio 0.05. IPO: 1 epoch, lr 5×10^{-6} , $\beta = 0.1$, batch 64, gradient accumulation 4. RLOO: 100 steps, lr 1×10^{-5} , batch 128, gradient accumulation 128, $G = 8$ rollouts per prompt, $\alpha_{\text{ent}} = \beta_{\text{KL}} = 0.001$, max log importance weight 5, sampling temperature 1.0 with no top- k /top- p filtering (per project spec).

SEC variants. See Table 1 for hyperparameters of v1 (faithful Chen et al. reproduction), v2 and v3 (sharpened bandit with bug fixes between them), and v4 (frontier-weighted). All SEC variants share the same RLOO base configuration above. For v4 we report 3 seeds at $\lambda = 1$ and 1 seed at $\lambda = 2$.

Variant	Bin key	T	α	Reward signal	UCB / floor	Notes
v1	n_{operands} (2 bins)	1.0	0.1	mean $ \text{adv} $	none	faithful SEC
v2	(n, q_{target}) (8 bins)	0.3	0.1	composite (normalized)	$c=1$, min 2	UCB bug; norm bug
v3	same	0.3	0.05	composite (raw)	$c=1$, min 4	bugs fixed
v4	same	0.3	0.05	frontier (Eq. 2)	$c=1$, min 4	$\lambda \in \{1, 2\}$

Table 1: SEC variants tried in this study. “UCB / floor” columns reflect the exploration bonus coefficient c and the min-visits floor.

Evaluation. We use the project-provided vLLM-based evaluator. 50 prompts on the held-out Countdown 3-to-4 test split, $K = 16$ samples per prompt. Sampling parameters: $T = 0.6$, top- $p = 0.95$, top- $k = 20$. Primary metric is the autograder’s average score (mean per-rollout reward); secondary metrics are pass@1 and pass@16 (fraction of prompts with at least one correct rollout in K samples), and fraction of rollouts that earn a full 1.0.

Baselines and choice of comparison. The relevant baseline for the SEC extension is RLOO with uniform prompt sampling. Comparing to SFT or IPO would be uninformative because SEC operates

only on the prompt distribution in the RL phase. We retain the SFT and IPO numbers as context to demonstrate that our underlying pipeline is working at the level expected by the project.

Compute. All training and evaluation runs use a single Modal H100 80GB instance. RLOO 100-step runs take ≈ 3.5 h; 300-step runs ≈ 9.5 h. Total wall-clock for the experiments reported in this paper is approximately 50 H100 hours.

5 Results

5.1 Pipeline baselines

Each pipeline stage clears its CS224R autograder threshold (Figure 6). The fixed RLOO baseline at 0.535 is the operative comparison point for all SEC variants.

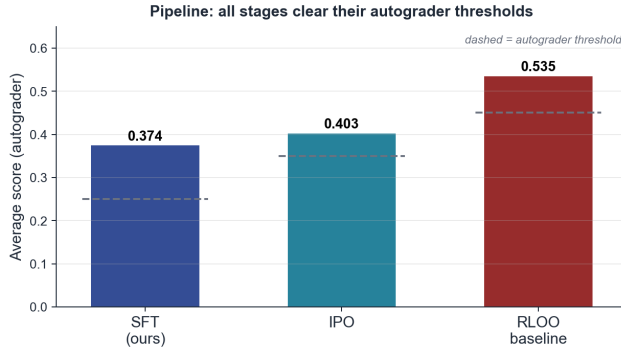


Figure 6: SFT, IPO, and RLOO baseline test scores. Dashed bars at thresholds (SFT 0.25 / IPO 0.35 / RLOO 0.45).

5.2 Quantitative: SEC variants vs. uniform baseline (100 steps)

Table 2 reports all SEC variants alongside the uniform baseline and pipeline references. No SEC variant exceeds the uniform baseline. Most differences fall inside the seed-to-seed standard deviation we measure for v4 ($\sigma \approx 0.015$).

Method	Avg	Pass@1	Pass@16	Frac correct
SFT (our retrain)	0.374	0.272	0.740	0.316
IPO	0.402	0.488	0.740	0.350
RLOO uniform baseline	0.535	0.546	0.760	0.514
SEC v1 (faithful)	0.478	0.562	0.720	0.451
SEC v2 (composite, w/ bugs)	0.486	0.372	0.680	0.450
SEC v3 (composite, fixed)	0.499	0.494	0.700	0.469
SEC v4 $\lambda=1$, seed 0	0.529	0.436	0.720	0.500
SEC v4 $\lambda=1$, seed 1	0.498	0.538	0.700	0.480
SEC v4 $\lambda=1$, seed 2	0.511	0.486	0.740	0.490
SEC v4 $\lambda=1$ mean \pm std (n=3)	0.513 \pm 0.015	0.487	0.720	0.490
SEC v4 $\lambda=2$ (single seed)	0.467	0.436	0.660	0.436

Table 2: 100-step test-set results across all SEC variants. All numbers are evaluated with the project’s vLLM evaluator using $T = 0.6$, $\text{top-}p = 0.95$, $\text{top-}k = 20$. The frontier-weighted v4 with three seeds is the most rigorous comparison. A one-sample t -test treating the baseline as a known constant gives $t = (0.513 - 0.535)/(0.015/\sqrt{3}) = -2.56$ on $df = 2$, which fails to reject the null of equality at $p = 0.05$ (critical $|t| \approx 4.30$). With only $n = 3$ seeds, our power to detect a ~ 0.02 effect against the seed-to-seed noise of ~ 0.02 is low; we report this as “we cannot demonstrate improvement and we cannot demonstrate equivalence.”

Training-time rollout reward (Figure 7) shows the same picture in continuous form: the baseline trajectory ends slightly higher than all SEC trajectories, but the seed-to-seed spread of v4 is comparable to the baseline–v4 gap.

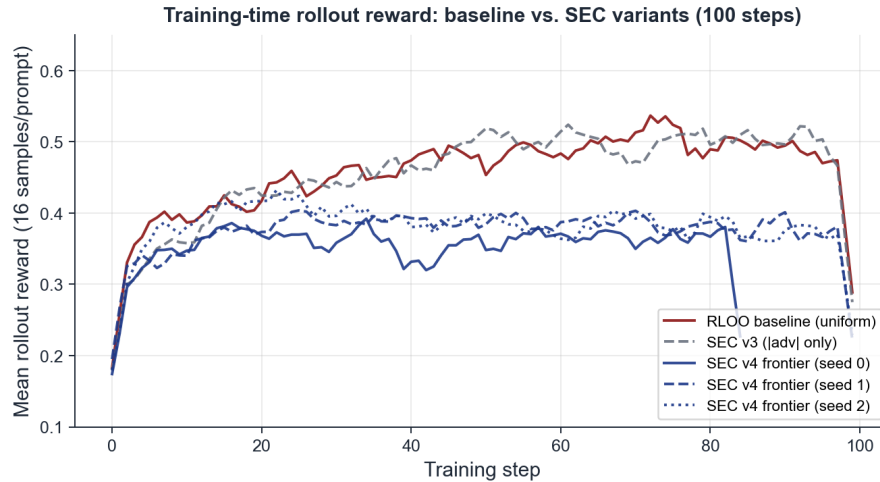


Figure 7: Mean rollout reward during 100-step training: baseline (uniform), SEC v3 (composite), and SEC v4 frontier across three seeds. Reward smoothed with a 5-step rolling mean for visualization.

5.3 Bandit redirection: the frontier signal does its job

Under the frontier-weighted signal, the bandit shifts its per-bin sampling toward harder bins (Figure 8). This is a direct measurement from per-bin visit counts logged to W&B, independent of test scores. Standard SEC v3 ends with 53% of total sampling on the easy (3-operand) bins versus 47% on the hard (4-operand) bins (a mild anti-curriculum). SEC v4 at $\lambda = 1$ samples 35% easy / 65% hard. The bandit’s Q-values move as expected (Figure 9): all 4-operand bins (red) accumulate higher Q than 3-operand bins (blue) under the frontier-weighted reward.

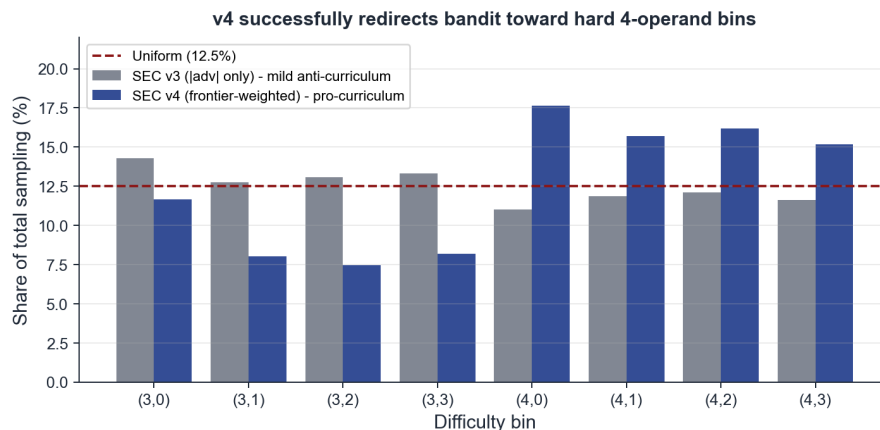


Figure 8: Total sampling share per bin (final 100-step value) for SEC v3 (gray) vs. SEC v4 $\lambda = 1$ (blue). Dashed red line is uniform (12.5% per bin).

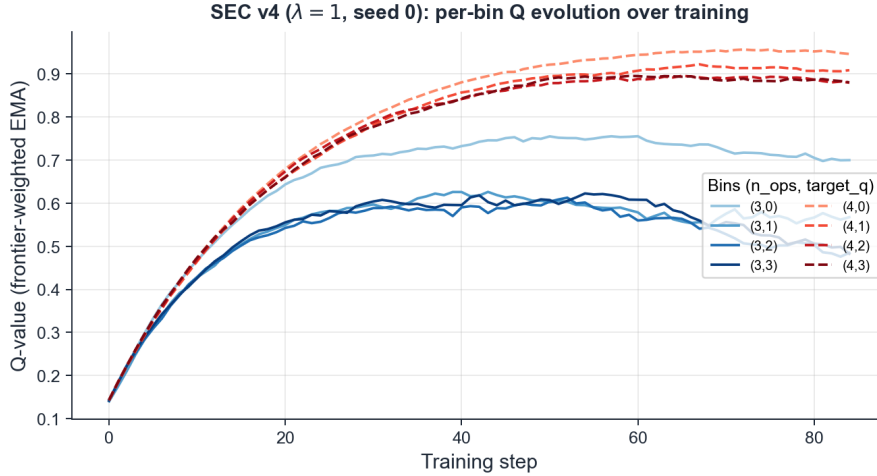


Figure 9: Per-bin Q-value evolution during SEC v4 ($\lambda = 1$, seed 0) training. 4-operand bins (red shades) accumulate higher Q-value than 3-operand bins (blue shades), reflecting the frontier signal’s directional intent.

5.4 Per-bin breakdown: where v4 helps and hurts

Per-bin test scores (Figure 10) show that v4 may help on the medium-hard bin (4, 1) where the model has some success but room to grow, and may consistently hurt the very hardest bin (4, 0) where the policy almost cannot solve the problem at all. The pattern is suggestive but does not survive seed noise rigorously: standard deviation across 3 seeds on the (4, 1) bin is 0.12, comparable to the apparent improvement. We mention this pattern as a candidate observation worth replicating with more seeds, not as a confirmed result.

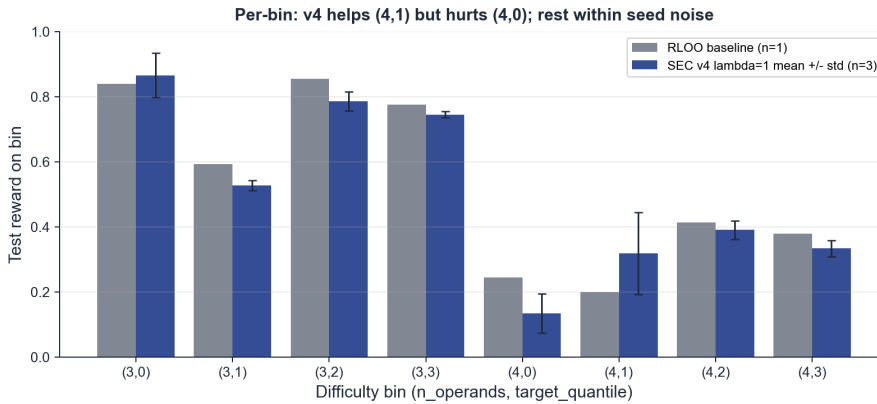


Figure 10: Per-bin test reward: uniform baseline (gray, $n = 1$) vs. SEC v4 $\lambda = 1$ mean \pm std ($n = 3$ seeds). The (4, 0) and (4, 1) bins are where the most interesting differences sit, but error bars cover most of these differences.

5.5 Qualitative analysis: where the model goes right and wrong

To give a qualitative sense of policy behavior, Listing 1 shows a correct test-set rollout from the SEC v4 model: the model briefly explores incorrect operations, then identifies the correct sequence. Listing 2 shows a format-only failure: the model emits a parseable `<answer>` tag (earning 0.1 reward) but the equation it produces evaluates to 82, not the target 72. The format-compliant-but-arithmetically-wrong category is common across all bins in our eval (Figure 4); contrary to our initial mechanism conjecture, it does *not* dominate the hardest bins (on bin (4, 0), 62.5% of rollouts produce no parseable answer at all, vs. 14.6% format-only and 22.9% correct).

```

<think>
Let me try to reach 57 using these numbers.
First, let me try working with 92 since it's the largest number:
92 - 50 = 42 (getting closer to 57)
42 + 4 = 46 (still not quite there)

Let me try one last time:
92 - 50 = 42
42 + 11 = 53
53 + 4 = 57 (This works!)
</think>
<answer>(92 - 50) + 11 + 4</answer>

```

Listing 1: SEC v4 correct test-set rollout. Target 57 from {4, 50, 92, 11}. Scored 1.0.

```

<think>
... (model tries many sequences, none give 72)
AHA! I see a possibility:
76 - 27 = 49
49 + 35 = 84
84 - 2 = 72 // arithmetic is wrong: 84 - 2 = 82, not 72
</think>
<answer>(76 - 27) + 35 - 2</answer>

```

Listing 2: SEC v4 format-only failure. Target 72 from {35, 2, 76, 27}. Model emits a parseable <answer> but the equation evaluates to 82, not 72. Scored 0.1.

This category of failure — the policy reliably emits format-conformant answers on hard problems but cannot solve the underlying math — is exactly the situation we predicted in §3.3 would compress $|\text{adv}|$. Across all 16 rollouts on the prompt in Listing 2, the policy never produces a correct answer, but most rollouts earn 0.1 for being parseable.

6 Discussion

Limitations. The limitations of this study are statistical and scope-related. We have three seeds for SEC v4 $\lambda = 1$ but only one for the uniform baseline and one for v4 $\lambda = 2$; a proper two-sample test would require matched seeds for the baseline. We cannot reject the null of v4 equality with the baseline at $p = 0.05$, but with $n = 3$ we also have low power to detect modest effects. The model size is held fixed at 0.5B and the budget at 100 steps (with one extended 300-step probe). Our initial mechanism conjecture for the $|\text{adv}|$ asymmetry — ternary verifier reward compressing within-group variance on hard bins by clustering rollouts at the 0.1 partial-credit value — was not supported by post-hoc analysis: the eval-time reward distribution on hard bins clusters at 0.0 (no answer), not at 0.1, and computing expected $|\text{adv}|$ from the eval distribution does not reproduce the training-time asymmetry (Figure 5). The asymmetry is therefore plausibly a training-time-specific phenomenon involving some combination of sampling temperature ($T = 1.0$ during training vs. $T = 0.6$ at eval), policy state mid-training, and reward structure; we cannot isolate any of these factors with our data. A strict binary-reward ablation (drop the 0.1 partial credit), matched-seed comparisons across training temperatures, and per-prompt $|\text{adv}|$ traces over training are the controlled experiments we would run to pin this down. Our literature review for novelty claims is limited; the curriculum learning field is large and we cannot rule out that the regularity we report has been characterized in prior work we have not surveyed.

Why the redirected curriculum does not improve test reward. We do not have a definitive answer; plausible candidates are: (a) at our budget the policy has too little training time to convert any curriculum advantage into measurable test performance; (b) the easier-bin sampling that v4 reduces was actually informative for the model’s general competence, and we paid that cost without offsetting it; or (c) our gating heuristic in Equation 2 is too coarse — the model may benefit from a bin only when its success rate lies in some intermediate range (say 5–20%), and gating only at $\bar{r} > 0$ treats a bin with 1% success rate the same as a bin with 50% success rate. Our 300-step probe is weakly consistent with (c): under uniform sampling, the policy reaches peak reward around step 50–80 then

drifts down, and v4 appears to hold harder-bin performance better than baseline while losing on mid-easy bins. Both runs are single-seed at 300 steps, so we cannot distinguish this pattern from seed noise.

Broader impact. The contribution we offer is methodological: an empirical observation about SEC’s reward signal in a previously uncharacterized regime, with a verified algorithmic intervention at the bandit level. We do not recommend deploying frontier-weighted SEC over uniform sampling at our scale. We do recommend that practitioners considering SEC for LLM RL fine-tuning measure per-bin $|\text{adv}|$ over training early on and check that it scales with the curriculum’s intended direction before committing to the standard SEC reward signal. The per-bin diagnostic is cheap to add and would have flagged the misalignment to us in a fraction of one RLOO step.

Difficulties encountered. The single most impactful difficulty was the pad-token contamination bug in §3.2, which silently degraded RLOO for the first several days of experiments before we traced it. Modal infrastructure flakes (Ray actor SIGSEGVs on container startup, gRPC heartbeat timeouts) caused ~ 5 run restarts but were not caused by our code. Generating clean per-seed comparisons required relaunching after each Modal outage and was the main reason we did not run baseline seeds: the marginal seed for v4 felt more informative than the marginal seed for the baseline, given the asymmetry of what we were trying to claim.

7 Conclusion

We replicated SEC for RLOO on Countdown 3-to-4 with Qwen2.5-0.5B and 100 training steps. In this regime, the standard SEC reward signal — mean absolute advantage — allocates curriculum effort backwards: it samples easier 3-operand bins more than harder 4-operand bins, the opposite of curriculum learning’s intent. The asymmetry is robust to bandit sampling direction (persists across runs that oversample easy and runs that oversample hard bins) and therefore is not an artifact of SEC’s own sampling. We initially attributed this to Countdown’s ternary verifier reward compressing within-group reward variance on hard bins, but our post-hoc per-bin reward distribution analysis does not support that specific mechanism, and the asymmetry largely vanishes at eval-time sampling settings, suggesting the cause is a training-time-specific interaction of policy state, sampling temperature, and reward structure that we cannot isolate from our data. We proposed a frontier-weighted reward signal that adds an explicit “room to grow” bonus, and demonstrated via per-bin bandit logs that this redirects the bandit toward harder bins as intended. Across three seeds, the redirected curriculum does not statistically improve test reward over uniform sampling. Our take-home message is that practitioners should not assume SEC’s $|\text{adv}|$ signal scales with intended curriculum direction — it is cheap to check via a per-bin diagnostic in early training, and when it does not scale correctly, the bandit’s allocation can be steered with a modified reward signal even though end-task gains may require larger models or longer budgets. The specific next steps we believe most likely to move the needle are (1) a learnability gate refining Equation 2 so that bins below some success-rate threshold (e.g., 5%) are suppressed (our current gate is effectively always active in our setup), (2) a binary-reward ablation that drops the 0.1 partial credit to test the reward-structure hypothesis, and (3) replication at larger model sizes and longer training budgets, matching the regime in which SEC was originally demonstrated.

8 Team Contributions

This is a solo project. All implementation, experiments, analysis, and writing were performed by Garrett Alarcon. The implementation consisted of filling in the project-provided starter code (the `NotImplementedError` blocks in the SFT, IPO, and RLOO trainers) and writing the SEC bandit, diagnosis analyses, and frontier-weighted reward signal on top. Relative to the original project proposal, the plan focused on implementing SEC and reporting curriculum gains; the contribution scope shifted toward characterizing why SEC underperforms in the small-model regime after the initial replication did not exceed the uniform baseline.

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, 2024.
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2024.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009.
- Xiaoyin Chen, Sheng Liang, Jiwen Wang, Jiarun Wang, Lucas Pian, Hao Lin, Yi Ren, Ben Wang, Sirui Xia, and Lijun Wu. Self-evolving curriculum for LLM reasoning, 2025.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention, 2023.
- Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. TinyZero: A simple, free, and reproducible implementation of reasoning RL. <https://github.com/Jiayi-Pan/TinyZero>, 2025.
- Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments, 2019.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.