ZRC: Zero-Resource Curriculum Reinforcement Learning for LLMs

Anonymous Author(s)
Affiliation

Abstract

Reinforcement learning (RL) is a central paradigm for enhancing the reasoning ability of large language models (LLMs), yet existing curricula often depend on external difficulty labels or handcrafted schedules, limiting scalability and robustness. We present Zero-Resource Curriculum (ZRC), a dynamic framework that removes such reliance by organizing training from intrinsic signals. ZRC estimates difficulty through empirical accuracy and self-verification failure, clusters tasks adaptively using DP-means, and employs Bayesian Online Change Point Detection (BOCPD) to trigger reclustering only under distributional shifts. A stabilization phase mitigates abrupt changes, while bandit-based sampling balances exploration and exploitation across clusters. Together, these components yield an adaptive and stable curriculum that enables efficient RL-based post-training for reasoning-centric LLMs.

Keywords: Curriculum learning, Reinforcement learning, Large language models, Bandit sampling, Change-point detection

1. Introduction

LLMs show strong reasoning ability, but multi-step reasoning remains challenging. RL can help, yet suffers from inefficiency and instability, motivating curriculum learning that guides models through tasks of increasing difficulty.

E2H Reasoner [1] arranges tasks from easy to hard but relies on external labels. AdaRFT [2] adapts thresholds from reward signals, which can be noisy, and SPaRFT [3] improves efficiency with clustering and bandits but still depends on predefined difficulty. These methods show the value of curricula yet remain limited by external supervision.

We propose **ZRC**, a dynamic framework for RL-based post-training without external labels. ZRC lets the model self-estimate difficulty via answer inconsistency and self-verification failure, clusters problems adaptively with DP-means, uses Bayesian Online Change Point Detection (BOCPD) to trigger reclustering only under distributional shifts, and applies multi-armed bandit sampling to balance exploration and exploitation.

Our contributions are twofold: ZRC is adaptive, adjusting to evolving competence through intrinsic difficulty estimation and clustering, and stable, using BOCPD and stabilization to prevent abrupt drift.

2. Related Work

Reinforcement learning (RL)–based post-training for large language models (LLMs) has become a promising direction to improve reasoning and decision-making capabilities. Among them, curriculum learning techniques help shape which problems the model sees when, and recent works explore both static and adaptive curricula.

2.1 Curriculum Learning for LLMs via RL

Early work such as *E2H Reasoner* arranges tasks from easy to hard and phases out trivial ones, showing both empirical and theoretical gains but requiring external labels or handcrafted schedules [1]. *AdaRFT* adapts task thresholds online using recent rewards, though noisy signals can distort training [2]. *SPaRFT* integrates clustering with bandit sampling to improve efficiency but still relies on predefined difficulty metrics [3]. More recently, *SEC* formulates curriculum allocation as a nonstationary bandit co-learned with policy updates, achieving better generalization on planning and reasoning tasks [4]. Overall, these methods highlight the value of curricula but remain constrained by external supervision or fixed heuristics.

[*] Corresponding author

Algorithm 1 Zero-Resource Curriculum (ZRC)

```
Require: Dataset $D = \{x_i\}_{i=1}^N$, Initial policy $\pi_{\theta_0}$, Encoder $f_{\text{sem}}$,
    Hyperparameters $(\lambda, \tau_{cp}, T_{\min}, E_{\text{stab}}, N, \beta_{\text{KL}})$
Ensure: Trained policy $\pi_{\theta^*}$
Initialize difficulty $\hat{d}_i$ (pass-rate + self-verification)
Construct embeddings $e_i = [\hat{s}_i \oplus \hat{d}_i]$
Run DP-means to form $\mathcal{C}_0$
Initialize BOCPD and bandit states
for $t = 1$ to $T$ do
    Update $\hat{d}_i$ for batch $B$; recompute $e_i$
    Incrementally update clusters $\mathcal{C}_t$ with DP-means
    Update success-rate streams $S_k(t)$ for each cluster
    Apply BOCPD to compute $p(cp_t)$
    if $p(cp_t) > \tau_{cp}$ and $t - t_{\text{last\_rc}} \geq T_{\min}$ then
        Trigger Soft/Hard reclustering; run stabilization ($E_{\text{stab}}$)
    end if
    Select cluster arm via Thompson/UCB bandit
    Update policy $\pi_\theta$ with policy optimization algorithm (PPO/GRPO)
end for
```

2.2 Clustering, Change-Point Detection

DP-means (Dirichlet Process means). DP-means is a hard clustering method derived from the small-variance limit of Dirichlet Process mixtures. It extends k-means with a penalty λK, creating a new cluster when points are far from all centroids, thus adaptively adjusting the number of clusters during training.

Bayesian Online Changepoint Detection (BOCPD). BOCPD maintains a posterior over run lengths. With conjugate models, it computes the probability of a change point at each step. Reclustering is triggered only when statistical evidence supports a distributional shift, controlling needless oscillation and preserving stability.

3. Zero-Resource Curriculum (ZRC)

We propose Zero-Resource Curriculum (ZRC), a dynamic curriculum learning framework for LLM post-training. The framework integrates four key modules: (1) self-difficulty estimation, (2) nonparametric clustering via DP-means, (3) selective reclustering with Bayesian Online Change Point Detection (BOCPD), and (4) stabilization and bandit-based curriculum sampling. Algorithm 1 summarizes the procedure.

3.1 Self-Difficulty Estimation

For each problem x_i, the policy generates N responses $\{y_{i,j}\}_{j=1}^N$ under high-temperature sampling ($T = 0.8$–1.0). We consider two intrinsic signals:

Empirical accuracy (verifier pass rate). Let $c_{i,j} \in \{0, 1\}$ denote the correctness of the j-th response, measured by a exact answer match score.

$$a_i = \frac{1}{N} \sum_{j=1}^{N} c_{i,j}. \tag{1}$$

Self-verification failure. Each response is re-evaluated by the model with a verification prompt, producing a binary verdict $\nu_{i,j} \in \{0, 1\}$.

$$v_i = \frac{1}{N} \sum_{j=1}^{N} \nu_{i,j}. \tag{2}$$

The final difficulty score increases when accuracy is low and self-verification fails often:

$$\hat{d}_i = \alpha \cdot (1 - a_i) + \beta \cdot v_i, \qquad \alpha + \beta = 1. \tag{3}$$

We then construct a joint embedding $e_i = [\, \hat{s}_i \oplus \hat{d}_i \,]$ where $\hat{s}_i = f_{\text{sem}}(x_i)$.

3.2 Nonparametric Clustering with DP-means

We adopt DP-means to cluster $\{e_i\}$, which removes the need to fix the number of clusters K in advance. The objective is

$$\mathcal{L} = \sum_i \|e_i - \mu_{z_i}\|_2^2 + \lambda K, \tag{4}$$

where μ_{z_i} is the assigned centroid. A new cluster is created if

$$\min_c \|e_i - \mu_c\|_2^2 > \lambda. \tag{5}$$

Clusters closer than $\lambda/2$ are merged, and clusters smaller than $n_{\min}$ are removed.

3.3 Change-Point Detection with BOCPD

Reclustering every step is expensive and destabilizing. We therefore employ BOCPD to selectively trigger updates. For each cluster C_k, we maintain a success-rate stream $S_k(t)$. Observations follow a Beta-Binomial model:

$$r_{k,t} \sim \text{Binom}(n_{k,t}, \theta_k), \quad \theta_k \sim \text{Beta}(\alpha, \beta). \tag{6}$$

BOCPD maintains the run-length posterior and computes the change-point probability $p(cp_t)$. Reclustering is triggered if

$$p(cp_t) > \tau_{cp}, \quad t - t_{\text{last\_rc}} \geq T_{\min}. \tag{7}$$

This mechanism suppresses false alarms, avoids thrashing, and reacts only to meaningful competence shifts, i.e., sustained changes in cluster success rates beyond transient noise.

3.4 Reclustering and Stabilization

We distinguish two modes:

- **Soft reclustering:** only reassigns membership, preserving centroids.
- **Hard reclustering:** reruns DP-means with new cluster creation and merging.

After reclustering, we insert a stabilization window of E_{stab} steps with:

$$P^{(s)} = \alpha^{(s)} P_{\text{old}} + (1 - \alpha^{(s)}) P_{\text{new}}, \tag{8}$$

$$\beta_{\text{KL}}^{(s)} = \beta_{\text{base}} \cdot \Big(c \cdot \rho^{s} \Big), \quad c{=}2, \ \rho{=}(c)^{-1/E_{\text{stab}}}, \tag{9}$$

We double the KL penalty after reclustering to tighten the trust region and prevent abrupt drift, then decay it back during stabilization.

3.5 Bandit-Based Curriculum and RL Update

Each cluster is treated as a bandit arm. Thompson sampling draws

$$\theta_k \sim \text{Beta}(\alpha_k, \beta_k), \quad k^* = \arg\max_k \theta_k, \tag{10}$$

while UCB1 computes

$$\text{UCB}_k = \bar{r}_k + \sqrt{\tfrac{2 \log t}{n_k}}. \tag{11}$$

Thompson sampling suits high-uncertainty settings by balancing exploration and exploitation, while UCB1 is more deterministic and effective when variance is moderate and tighter regret bounds are desired.

The selected cluster supplies training data, and the policy is then updated with algorithms such as PPO or GRPO.

4. Conclusion

We introduced Zero-Resource Curriculum (ZRC), a framework for RL-based post-training of LLMs that removes reliance on external difficulty labels. ZRC integrates self-difficulty estimation, adaptive clustering, change-point detection, and bandit sampling with stabilization to provide an adaptive curriculum for reasoning tasks.

Reference

[1] J. Kim and o. Han, “Curriculum reinforcement learning from easy to hard tasks improves llm reasoning,” *Proceedings of the 2025 International Conference on Learning Representations (ICLR)*, 2025, arXiv:2506.06632.

[2] Xu and collaborators, “Efficient reinforcement finetuning via adaptive curriculum learning,” *Proceedings of the 2025 International Conference on Learning Representations (ICLR)*, 2025, arXiv:2504.05520.

[3] Chen and collaborators, “Sparft: Self-paced reinforcement fine-tuning for large language models,” *Proceedings of the 2025 International Conference on Machine Learning (ICML)*, 2025, arXiv:2502.xxxxx.

[4] Zhou and collaborators, “Self-evolving curriculum for llm reasoning,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2025, arXiv:2505.14970.