
RLEP: Reinforcement Learning with Experience Replay for LLM Reasoning

Hongzhi Zhang, Jia Fu, Jingyuan Zhang, Kai Fu, Qi Wang, Fuzheng Zhang, Guorui Zhou

Klear Team, Kuaishou Technology

Abstract

Reinforcement learning (RL) for large language models is an energy-intensive endeavor: training can be unstable, and the policy may gradually drift away from its pretrained weights. We present *RLEP*—Reinforcement Learning with Experience rePlay—a two-phase framework that first collects verified trajectories and then replays them during subsequent training. At every update step, the policy is optimized on mini-batches that blend newly generated rollouts with these replayed successes. By replaying high-quality examples, RLEP steers the model away from fruitless exploration, focuses learning on promising reasoning paths, and delivers both faster convergence and stronger final performance. On the Qwen2.5-Math-7B base model, RLEP reaches baseline peak accuracy with substantially fewer updates and ultimately surpasses it, improving accuracy on AIME-2024 from 38.2% to 39.9%, on AIME-2025 from 19.8% to 22.3%, and on AMC-2023 from 77.0% to 82.2%. Our code, datasets, and checkpoints are publicly available at <https://github.com/Kwai-Klear/RLEP> to facilitate reproducibility and further research.

1. Introduction

Large language models (LLMs) have recently made rapid progress in reasoning. OpenAI o1 (OpenAI, 2024), DeepSeek R1 (Guo et al., 2025), and Qwen3 (Yang et al., 2025) etc. has established a new paradigm for solving complex problems. Reinforcement learning (RL) is a key driver of this advance: R1 shows that even a simple rule-based reward can consistently improving reasoning ability.

Reinforcement Learning with Verifiable Rewards (RLVR) is no trivial task, it must simultaneously balance learning capacity, policy stability, and exploration ability. **Learning capacity** ensures that policy updates absorb the knowledge uncovered during exploration; **policy stability** keeps gradients within reasonable bounds and have the LLM’s weights remain close to their pre-training initialization so as to avoid catastrophic forgetting; and **exploration ability** allows the model to discover informative trajectories that make continued training worthwhile. Recent work has pinpointed techniques for achieving this balance: DAPO (Yu et al., 2025) and DrGRPO (Liu et al., 2025) introduce a token-mean objective that strengthens learning over long sequences, *clip-higher* biases the model toward low-probability positive rewards to prevent collapse of the exploration space, and adopting a high-entropy token-updating strategy (Wang et al., 2025) is proposed to improve both efficiency and stability.

Nevertheless, RL training is an energy-intensive journey. A policy departs from its current parameters, explores reward-bearing reasoning paths, and incorporates what it learns. As training continues, training instability and weight drift can cause progress to plateau—or even regress—so that most runs eventually saturate at a fixed performance level.

The training process is akin to a mountaineering expedition. As Figure 1 shows, the climber first scouts four possible routes, ultimately reaching the initial red flag—the highest point attainable before exhaustion sets in. On the next journey, the climber retraces those marked waypoints, conserves energy, and pushes onward to a second, higher flag. Motivated by this analogy, we introduce Reinforcement Learning with Experience rePlay (RLEP). By replaying successful trajectories from earlier runs, the policy rapidly recovers its previous best performance at minimal cost and then surpasses it with stable, continued gains.

RLEP comprises two phases—experience collection and replay-based training. **Experience collection.** Starting from a model trained with vanilla RL, we generate multiple reasoning trajectories for every question. Trajectories that arrive at the correct answer are retained, forming an experience pool of verified solutions. **Replay-based training.** During each training step, the policy is updated with a mix of freshly generated rollouts and a small batch of successful trajectories sampled from the experience pool. This blend lets the model reinforce proven reasoning chains while still exploring new ones, yielding faster convergence and stronger final performance.

All experiments were performed with the Qwen-2.5-7B model. Experience replay allows the policy to reclaim its previous best score in a few updates and then exceed it latter. Consequently, RLEP achieves stable accuracy gains over the baseline—rising from 38.2% to 39.9% on AIME 2024 (+1.7pp), from 19.8% to 22.3% on AIME 2025 (+2.5pp), and from 77.0% to 82.2% on AMC 2023 (+5.2pp). These results show that experience replay not only accelerates convergence but also produces a higher final performance ceiling.

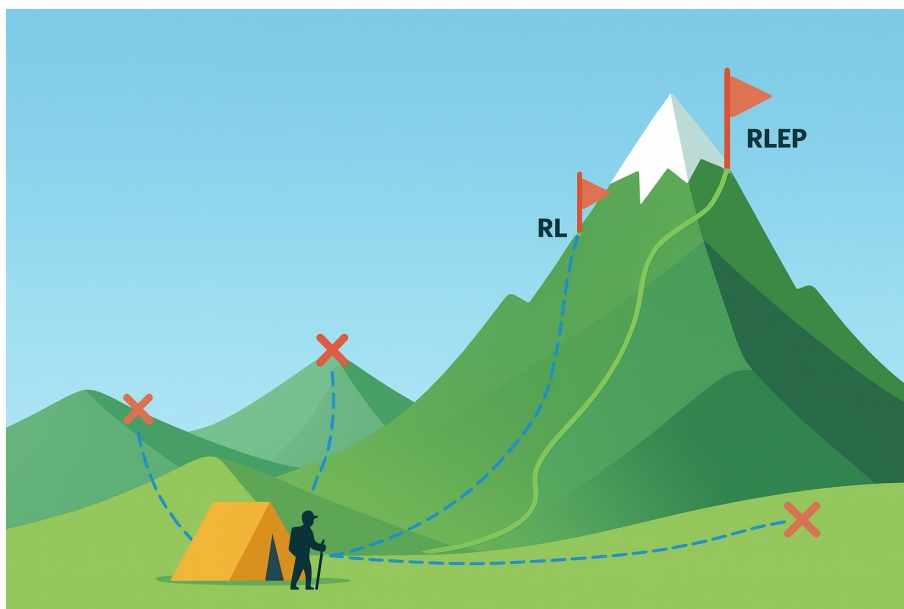


Figure 1 | Illustration of reinforcement learning with experience replay. During the first trip, the climber explores four candidate routes and reaches the initial red flag, but must stop there due to limited energy. With RLEP, the climber quickly replays the successful trajectory to the first flag and then ascends farther to a higher peak.

2. Related Work

Experience Replay is a well-established technique in reinforcement learning that significantly improves sample efficiency and stabilizes training (Lin, 1992; Mnih et al., 2015; Schaul et al., 2015). Lin (1992) introduced the original ER framework; Mnih et al. (2015) integrated it into DeepQ-Networks (DQN) and demonstrated its pivotal role in deep RL; and Schaul et al. (2015) proposed Prioritised Experience Replay to further enhance sampling efficiency.

Experience Replay (ER) has recently gained traction in RL training of large language models. Existing studies typically deploy replay within the same training run to salvage hard prompts—those for which the current policy cannot yet generate a correct rollout. EFRAME (Wang et al., 2025), for example, performs extra rollouts on such difficult cases and replays only the trajectories judged valuable, yielding consistent improvements on multimodal tasks. The Rollout-Rescue Mechanism (An et al.) adopts a simpler strategy: whenever training encounters a failure, it randomly replaces one incorrect response with a previously buffered correct answer from an earlier epoch. LUFFY (Yan et al., 2025) instead leverages powerful offline guidance like DeepSeek R1 for prompts lacking correct on-policy rollouts and redesigns advantage estimation to remain valid despite the large policy gap. Dou et al. (2025) approaches the issue from another angle, using the PPO critic to identify promising states explored early in training and replaying them to preserve exploration capacity. Moving beyond these hard-sample tactics, Bartoldson et al. (2025) proposes a replay-buffer-centred framework that decouples rollout generation from actor updates, simultaneously boosting decoding diversity and training speed.

In contrast, our method gathers trajectories from a converged policy—thereby drawing on inherently stable states—then restarts training from scratch while interleaving these stable trajectories with freshly sampled rollouts. The replay accelerates convergence and smooths learning, whereas the new rollouts safeguard exploration. Crucially, we apply ER uniformly to all prompts rather than restricting it to difficult cases, extending the benefits of replay to the entire training distribution.

3. RL with Experience Replay

3.1. GRPO: Methodology and Recent Advances

Review of GRPO. Following Shao et al. (2024) GRPO treats each question q by sampling a group of G candidate answers, $O = \{o_1, o_2, \dots, o_G\}$, from the current policy π_θ . A scalar reward r_i is assigned to every answer o_i , yielding a reward vector $\mathcal{R} = \{r_1, r_2, \dots, r_G\}$. The advantage is then used to scale the policy-gradient update, so that trajectories whose rewards stand above the group mean are reinforced, whereas those below the mean are suppressed:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (1)$$

Let π_θ be the current policy and $\pi_{\theta_{\text{old}}}$ the behaviour policy from the previous iteration. Omitting the KL-divergence regulariser, GRPO updates θ by maximising the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)} \left[\frac{1}{G} \sum_{i=1}^G \left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon\right) A_i\right) \right) \right] \quad (2)$$

Enhancements: *token-mean* and *clip-higher*. Two refinements have been proposed to stabilise GRPO training:

- (a) **Token-mean** Liu et al. (2025); Yu et al. (2025). Instead of taking a *sequence*-level average and then performing a macro-average over the G answers, token-mean averages the log-probability ratios *token by token*. This prevents long, erroneous sequences from being under-penalised and preserves the learning signal for long, correct sequences.
- (b) **Clip-higher** Yu et al. (2025). Positive-advantage trajectories are clipped with a *higher* upper bound than the standard PPO limit, while negative-advantage trajectories keep the usual lower bound. This asymmetric clipping mitigates entropy collapse during RL training and plays a key role in balancing exploitation with continued exploration.

With these two strategies, the GRPO objective is revised to

$$\mathcal{J}_{\text{GRPO}+}(\theta) = \mathbb{E}_{\substack{q \sim P(Q), \\ \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)}} \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) A_{i,t}, \text{clip} \left(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}} \right) A_{i,t} \right) \right] \quad (3)$$

3.2. RL training with Experience Replay

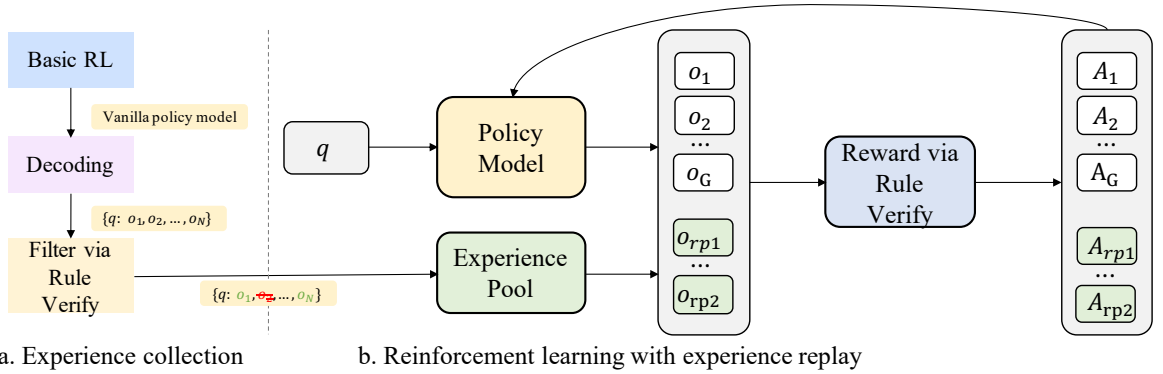


Figure 2 | **RLEP training pipeline.** **(a) Experience collection.** After a preliminary vanilla RL run, the seed policy decodes multiple trajectories for each problem; those that reach a verified correct answer are retained and stored in an experience pool. **(b) Replay training.** At every update the current policy rolls out G fresh trajectories (blue). We then sample M successful trajectories from the experience pool (green) and merge them, yielding an enlarged batch of $G' = G + M$. Advantages are computed over this mixed set.

The complete RLEP workflow—comprising the two stages of *experience collection* and *replay training*—is illustrated in Figure 2.

Experience Collection As illustrated in Figure 2a, we begin with a conventional RL run to obtain a seed policy. This policy decodes multiple candidate trajectories for every question, and any trajectory that reaches a verified correct answer is retained and stored in an *experience pool* for later replay. For every question we maintain an *experience pool*—a set of trajectories that reach the correct answer, collected from previous RL-trained checkpoints.

RLEP Training. Figure 2.b gives an overview of the RLEP training workflow. During each training update, we mix fresh roll-outs with replayed successes as follows.

- (i) **Rollout.** The current policy π_θ generates a group of G candidate trajectories, as in standard GRPO.
- (ii) **Experience replay.** We randomly sample M successful trajectories from the experience pool and append them to the freshly generated rollouts, expanding the group size from G to $G' = G + M$.
- (iii) **Policy update.** Given the enlarged set of trajectories $\{o_i\}_{i=1}^{G'}$ ($G' = G + M$), we apply the same token-level, asymmetrically clipped GRPO rule, now computed over the *mixed* group of fresh roll-outs and replayed successes:

$$\mathcal{J}_{\text{RLEP}}(\theta) = \mathbb{E}_{\substack{q \sim P(Q), \\ \{o_i\}_{i=1}^{G'} \sim \pi_{\theta_{\text{old}}}(O|q)}} \left[\frac{1}{\sum_{i=1}^{G'} |o_i|} \sum_{i=1}^{G'} \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) A_{i,t}, \text{clip} \left(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}} \right) A_{i,t} \right) \right] \quad (4)$$

where $r_{i,t}(\theta)$ is the token-wise importance ratio $\pi_\theta(o_{i,t} | q) / \pi_{\theta_{\text{old}}}(o_{i,t} | q)$. The advantage term $A_{i,t}$ is standardised over *all* G' trajectories so that replayed successes and new roll-outs share a common baseline:

$$A_{i,t} = \frac{r_{i,t} - \text{mean}(\{r_{1,t}, \dots, r_{G',t}\})}{\text{std}(\{r_{1,t}, \dots, r_{G',t}\})}. \quad (5)$$

Equations (4) and (5) thus generalise the GRPO update to the RLEP setting, ensuring that every update step jointly leverages fresh roll-outs and (replayed high-quality trajectories).

Replaying curated successful trajectories shields the policy from unproductive exploration, concentrates learning on fruitful lines of reasoning, and ultimately yields both quicker convergence and the potential to higher final accuracy.

4. Experiment

4.1. The Optimized Baseline

Training strategy Starting from the hyper-parameter recommendations in DAPO, we make several light adjustments and obtain consistent gains on AIME-2024, AIME-2025 and other datasets. Specifically, we adopt the *token-mean*, *clip-higher* and *overlong-reward-shaping* strategies, while keeping most of the default settings in Verl (Sheng et al., 2024). Considering rollouts dominate the wall-clock cost, we intentionally omit the *dynamic-sample* acceleration scheme. Instead, we fine-tune the remaining hyper-parameters to build a stronger baseline.

With 512 samples per rollout, the original configuration performs 16 actor updates using 32-sample mini-batches. Although this setting converges quickly, we observe a late-stage decline in both BoN and Maj@N accuracy. Increasing the mini-batch size to 64—i.e., eight updates per rollout—substantially improves training stability.

Implement details. We utilize Qwen2.5-Math-7B (Yang et al., 2024) as the base model, the input context length is set to 1024 while the maximum response length is set to 3072. Each rollout

Method	<i>token-mean, clip-higher, overlong-shaping</i>	<i>dynamic-sampling</i>	ppo_mini_batch_size
DAPO	✓	✓	32
DAPO-nodyn-bs32	✓	✗	32
DAPO-nodyn-bs64	✓	✗	64

Table 1 | The tuned baseline configurations based on DAPO.

stage takes 512 prompts. The temperature and top- p value are set to 1.0 during training and validation.

Figure 3 compares several variants:

- DAPO vs. DAPO-nodyn-bs32. We find that DAPO with dynamic sampling achieves higher accuracy, show that the positive impact of dynamic sampling.
- Comparing DAPO-nodyn with different PPO training mini-batchsize, the 32-sample mini-batch learns faster at the beginning, but the 64-sample mini-batch ultimately converges to higher accuracy and a smoother Maj@32 curve. The DAPO-nodyn-bs64 even slightly surpasses DAPO in overall accuracy, eliminating the impact of removing of dynamic sample. In practice, each DAPO update takes roughly 220s before step 230, whereas DAPO-nodyn-bs64 needs only about 160s. After step 230, DAPO’s per-step time climbs to around 360s because additional rollouts are required to fill the batch. Balancing speed and accuracy, we therefore use the configure DAPO-nodyn-bs64 for subsequent RLEP experiments.
- Model accuracy climbs quickly at the start of training, but as instability accumulates and the policy drifts from its initial weights, overall accuracy eventually plateaus—and can even decline—after a certain number of steps. This phenomenon shows that RL training is an energy-intensive journey.

Typically, batch size has only a modest effect in standard supervised fine-tuning (SFT). In reinforcement learning, each rollout is followed by several policy-update steps, and batch size directly influences the proportion of samples affected by the advantage clipping operation. This coupling appears to be the reason batch size matters much more in RL.

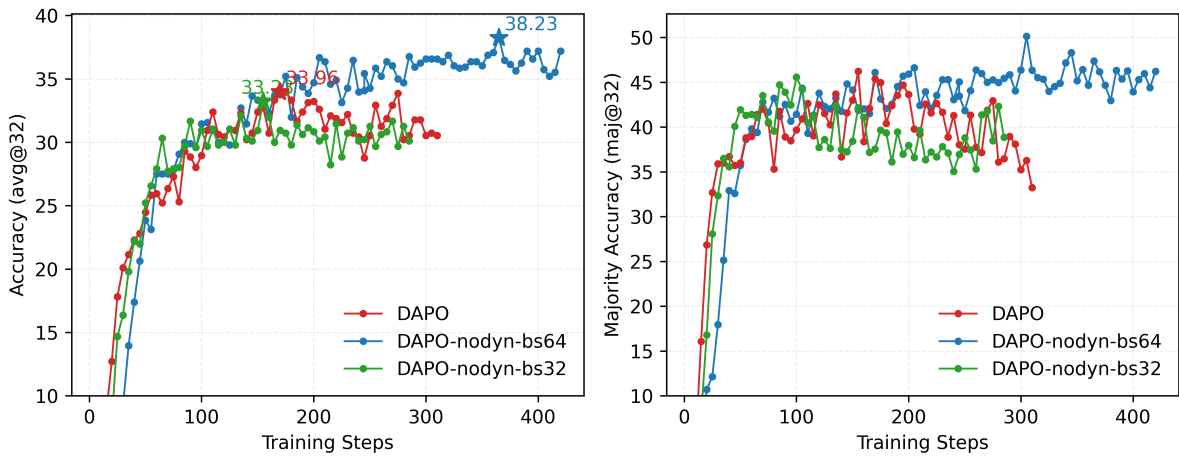


Figure 3 | Performance of the optimized baseline method.

4.2. RLEP Experimental Results

We start from the DAPO-nodyn-bs64 baseline, which trains for 400 PPO steps with a mini-batch size of 64 to build the experience pool. For every question, the policy samples 64 candidate answers (temperature 0.7, top-p 0.95); only answers verified as correct by the reward model are kept, and we require at least two such valid reasoning paths per question. During the RLEP phase, each question receives 16 fresh on-policy rollouts plus 2 replayed answers, while all other hyper-parameters remain identical to the baseline. Per-step runtime increases by under 5s relative to the DAPO-nodyn-bs64 baseline, leaving overall training time essentially unchanged.

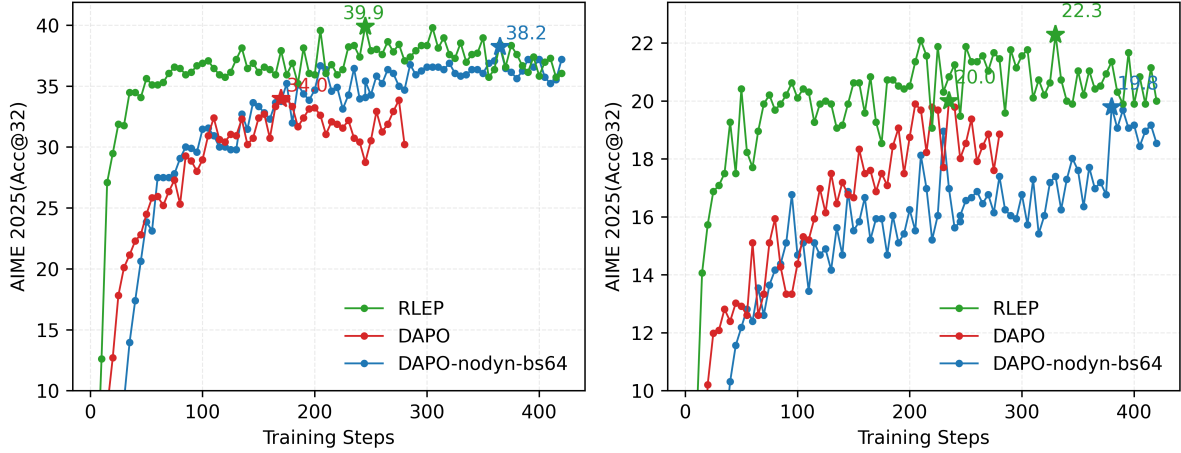


Figure 4 | Main experimental results of RL with Experience Replay.

The main results are shown in Figure 4 and can be summarised as follows:

- **Rapid early gains.** With replayed experience, accuracy rises sharply at the start of training. On the AIME-2024 dataset, RLEP matches the baseline’s peak performance by step 135 (the baseline needs 380 steps). On AIME-2025, it surpasses the baseline’s best score after only 50 steps. The replayed trajectories steer the model away from unproductive early exploration and difficult reasoning paths.
- **Higher final performance.** RLEP does more than accelerate convergence—it finishes higher. The best accuracy on AIME-2024 improves from 38.2% to 39.9%, and on AIME-2025 from 19.8% to 22.3%. (These points are peak value of the line, but similar conclusion could be drawn from the whole training procedure.) Evaluated offline on the unseen AMC-2023 dataset, accuracy rises from 77.0% to 82.2%. These results show that leveraging prior experience enables RLEP to converge to superior solutions.

Additionally, we examined whether supplementing the replay buffer with failed answers could help the policy avoid poor solutions. Replaying both successful and unsuccessful trajectories, however, produced no measurable improvement over positive replay alone. Error patterns vary widely across models and training stages—the mistake space is simply too broad—so unlikelihood updates on these heterogeneous errors provide little benefit to the current policy.

5. Conclusion and Future Work

In this technical report we introduce Reinforcement Learning with Experience rePlay (RLEP). During the experience-collection phase, earlier RL runs act as pathfinders, tracing high-reward

trajectories. In the subsequent replay stage, the model rapidly converges to these previously discovered optima and then pushes beyond them to reach even stronger performance. Experimental results on AIME 2024, AIME 2025 and AMC 2023 demonstrates the effectiveness of RLEP.

In future work, we will further explore: (1) devising smarter experience-selection schemes that leverage offline heuristics and model-based rewards to identify the most informative reasoning paths for replay, and (2) extending RLEP beyond the single-dataset setting by training on much larger corpora and evaluating its effectiveness across different domains.

References

- [1] C. An, Z. Xie, X. Li, L. Li, J. Zhang, S. Gong, M. Zhong, J. Xu, X. Qiu, M. Wang, and L. Kong. Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models. URL <https://hkunlp.github.io/blog/2025/Polaris>.
- [2] B. R. Bartoldson, S. Venkatraman, J. Diffenderfer, M. Jain, T. Ben-Nun, S. Lee, M. Kim, J. Obando-Ceron, Y. Bengio, and B. Kailkhura. Trajectory balance with asynchrony: Decoupling exploration and learning for fast, scalable llm post-training, 2025. URL <https://arxiv.org/abs/2503.18929>.
- [3] S. Dou, M. Wu, J. Xu, R. Zheng, T. Gui, Q. Zhang, and X. Huang. Improving rl exploration for llm reasoning through retrospective replay, 2025. URL <https://arxiv.org/abs/2504.14363>.
- [4] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [5] L.-J. Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Mach. Learn.*, 8(3–4):293–321, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992699. URL <https://doi.org/10.1007/BF00992699>.
- [6] Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.
- [7] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. URL <https://api.semanticscholar.org/CorpusID:205242740>.
- [8] OpenAI. Openai o1 system card. <https://openai.com/index/openai-o1-system-card/>, Dec. 2024. Updated December 5, 2024.
- [9] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.
- [10] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [11] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

- [12] C. Wang, L. Wei, Y. Zhang, C. Shao, Z. Dan, W. Huang, Y. Wang, and Y. Zhang. Eframe: Deeper reasoning via exploration-filter-replay reinforcement learning framework, 2025. URL <https://arxiv.org/abs/2506.22200>.
- [13] S. Wang, L. Yu, C. Gao, C. Zheng, S. Liu, R. Lu, K. Dang, X. Chen, J. Yang, Z. Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.
- [14] J. Yan, Y. Li, Z. Hu, Z. Wang, G. Cui, X. Qu, Y. Cheng, and Y. Zhang. Learning to reason under off-policy guidance, 2025. URL <https://arxiv.org/abs/2504.14945>.
- [15] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, K. Lu, M. Xue, R. Lin, T. Liu, X. Ren, and Z. Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.
- [16] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [17] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.