Beyond Markovian: Reflective Exploration via Bayes-Adaptive RL for LLM Reasoning

Anonymous Authors[1]

Abstract

Large Language Models (LLMs) trained via Reinforcement Learning (RL) have exhibited strong reasoning capabilities and emergent reflective behaviors, such as backtracking and error correction. However, conventional Markovian RL confines exploration to the training phase to learn an optimal deterministic policy and depends on the history contexts only through the current state. Therefore, it remains unclear whether reflective reasoning will emerge during Markovian RL training, or why they are beneficial at test time. To remedy this, we recast reflective exploration within the Bayes-Adaptive RL framework, which explicitly optimizes the expected return under a posterior distribution over Markov decision processes. This Bayesian formulation inherently incentivizes both reward-maximizing exploitation and information-gathering exploration via belief updates. Our resulting algorithm, BARL, instructs the LLM to stitch and switch strategies based on the observed outcomes, offering principled guidance on when and how the model should reflectively explore. Empirical results on both synthetic and mathematical reasoning tasks demonstrate that BARL outperforms standard Markovian RL approaches, achieving superior test-time performance and token efficiency.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive reasoning abilities, such as in solving complex math problems. A key factor driving this progress is the use of Chain-of-Thought (CoT) reasoning (Wei et al., 2022), where the model engages in intermediate deliberation before producing an answer. Building on this, recent advances have employed Reinforcement Learning (RL) to further enhance LLM reasoning by optimizing for verifiable outcome rewards (Jaech et al., 2024; Guo et al., 2025; Yu et al., 2025; Wei et al., 2025). Notably, RL-trained models have exhibited emergent behaviors such as generating long CoTs and engaging in self-reflection, a process of backtracking to previous states to correct earlier mistakes, also known as the "Aha moment" (Guo et al., 2025; Zeng et al., 2025). However, despite these compelling phenomena, it remains unclear why and under what conditions reflective reasoning is beneficial at test time, or whether such behaviors will emerge through conventional RL training.

Prevalent views attempt to explain the usefulness of test-time reflections as exploratory steps that provide additional contexts for more optimal decision-making. Yet in standard Markovian RL, the exploration–exploitation trade-off is resolved entirely during training: the agent interleaves exploration and exploitation to learn a training-time optimal policy, but switches to pure exploitation at test time. As a result, standard RL allows a Markovian policy to be optimal by simply memorizing training solutions once encountered. Moreover, the Markov assumption restricts the policy to condition decisions solely on the current state rather than on contextual information gathered through exploration. Thus, the agent learns an optimal deterministic policy through repeated trial and error, with no incentives to adaptively explore with reflections. In summary, under conventional Markovian RL, there is no guarantee that reflective explorations will emerge during training, nor does it explain why such explorations might be advantageous at test time.

To address this gap, we propose grounding reflective reasoning with Bayes-Adaptive RL, which explicitly optimizes for test-time generalization by maximizing the expected return under a posterior distribution over MDPs. The objective incentivizes both reward-seeking actions and epistemic explorations that gather information to reduce the MDP's uncertainty, such as the uncertainty regarding the progress made by different actions. This enables the model to adapt on-the-fly at test time by updating its beliefs and switching strategies based on observed outcomes, naturally giving rise to reflective exploration behaviors. We prove that the expected return of an adaptive policy can be exponentially higher than the optimal Markovian policy at test time.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Building upon this formulation, we introduce a novel algorithm, *Bayes-Adaptive RL for LLM Reasoning* (BARL). For each prompt, BARL performs online rollouts to generate a set of candidate answers, each associated with an MDP hypothesis. The state-action value is then computed by weighting each hypothesis according to the model's current belief, with penalties applied for mismatches between predicted and observed rewards, thereby signaling when to switch strategies. BARL provides a principled mechanism for integrating and revising plausible strategies, analogous to linearizing best-of-N reasoning, but with explicit guidance on *when* and *how* the model should reflectively explore.

To illustrate the benefits of BARL, we begin with a synthetic task designed to mirror test-time generalization in LLM reasoning. The agent receives a reward only when it repeats a prompt token three times, but the training and testing prompt tokens differ. Markovian RL memorizes the training solutions and fails to generalize. In contrast, BARL learns to switch strategies by eliminating hypotheses, ultimately discovering the ground-truth MDP for optimal behavior.

We further evaluate BARL on math reasoning tasks using various LLMs. Across these models, BARL consistently outperforms Markovian RL algorithms, such as GRPO and a strong progress-reward baseline, on multiple benchmarks. BARL achieves significantly greater token efficiency, requiring up to **2x** fewer than GRPO and over **10x** fewer than the Qwen2.5-Math-1.5B base model. Moreover, we observe no strong correlation between overall model performance and the frequency of reflections. Instead, BARL's advantage stems from more efficient exploration and more effective thinking tokens. We summarize the takeaways as follows:

Key Takeaways: Why, How, and When to Reflect

- **Why:** Markovian RL neither ensures the emergence of reflective exploration nor explains its benefits at test time since (1) exploration is confined to training to learn an optimal deterministic policy, and (2) the state-conditional policy lacks incentives to collect additional contexts and backtrack. In contrast, Bayesian RL, by optimizing test-time generalizability, encourages explorations to gather contextual information that reduces the MDP uncertainty.
- **How:** BARL stitches plausible strategies by maintaining a posterior over MDP hypotheses, each associated with a candidate answer. Reflective exploration emerges through hypothesis elimination, enabling on-the-fly adaptation.
- **When:** LLMs should self-reflect when discrepancies arise between their internal beliefs and cumulative reward feedback—signaling strategy switching by downweighting hypotheses that are unlikely to be optimal given previous observations.

2 The Necessity of Bayes-Adaptive RL for Reflective Reasoning

Markovian RL. When the underlying MDP is known with certainty, the Markov property ensures that the policy and value depend on the history $h_t = (s_0, a_0, r_0, \ldots, s_t)$ only through the state s_t, i.e., $Q^\pi(h_t, a_t) = Q^\pi(s_t, a_t)$. In this setting, exploratory actions that aim to enrich the history h_t with additional contexts, such as incorrect attempts followed by backtracking, are unnecessary, as the current state s_t already encodes all relevant information for optimal decision-making.

Moreover, the optimal Q-function is $Q^* = \max_\pi Q^\pi$ and the optimal policy π^* is greedy w.r.t. Q^*. That is, π^* is a deterministic policy, where $\pi^*(a'|s) = 1$ for $a' = \operatorname{argmax}_a Q^*(s, a)$.

Theorem 2.1. Optimality of Markovian RL is attained by deterministic non-reflective policies.

Reflective policies are more suboptimal than non-reflective policies in both discounted infinite-horizon and finite-horizon MDPs, as the Q-value of the wrong action is no larger than that of the correct action since more tokens are needed to correct the error. Explorations should only occur during training, in a trial-and-error manner with repeated episodes, to discover the golden answers. The Markovian RL objective allows the optimal policy that memorizes these training answers to be fully exploited, with no incentive to adaptively explore with reflections.

For the non-standard undiscounted infinite-horizon MDPs, where LLMs are encouraged to generate *infinite* tokens *without concerning token efficiency*, reflective policies may be as optimal as non-reflective ones. This is because the Q-value of the wrong action *can* match that of the correct one if the error is eventually corrected through reflection. This observation provides a partial explanation for the emergence of "Aha moment" with long CoT. However, even in such settings, reflective reasoning may still fail to emerge under Markovian RL, particularly if golden answers are discovered either directly or by pruning incorrect exploratory steps. In other words, this only explains why reflective explorations *can* appear during Markovian RL, instead of why these behaviors are preferable to simply memorizing training solutions, nor whether they will emerge during training.

Next, we present Bayes-Adaptive RL, which explicitly optimizes for test-time generalization and naturally induces reflective explorations.

Bayes-Adaptive RL. In a Bayes-Adaptive MDP (BAMDP) (Bellman & Kalaba, 1959; Martin, 1965; Lidayan et al.), the agent maintains uncertainty over the underlying MDP, which is gradually reduced through interactions. Due to this implicit partial observability (Duff,

2001; Ghosh et al., 2021), the policy and value depend on the full history h_t, instead of only the state s_t, to capture the agent's evolving belief about the MDP parameters through cumulative observations. The objective for BAMDPs is

$$\mathcal{J}_{\text{Bayes}}(\pi_\theta) := \mathbb{E}_{s_0,\pi_\theta}\left[\sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{M}\sim p(\mathcal{M}|h_t)}\left[r_{\mathcal{M}}(s_t,a_t)\right]\right],$$

where $p(\mathcal{M}|h_t)$ is the posterior distribution of $\mathcal{M}$ after observing h_t. This objective encourages the agent to not only maximize immediate rewards but also explore to gather more context about the uncertain MDP.

Optimal adaptive policies naturally induce exploratory reflection behaviors, which provide additional contextual information even if the state remains identical. While reflective actions may be suboptimal relative to the (unknown) ground-truth MDP, the gathered context, especially the rewards, reduces the MDP's uncertainty. This enables future policies to leverage the updated belief to act more optimally.

Theorem 2.2. The test-time expected return of a Bayes-Adaptive policy can be *exponentially higher* in T^* than that of the optimal Markovian policy, where T^* is the minimal number of steps required to reach the correct answer under an optimal deterministic policy.

3 Method

The policy gradient for Bayes-Adaptive RL is as follows, which differs from (B.2) by replacing the value under a predefined $\mathcal{M}$ with a posterior-weighted value:

$$\nabla_\theta \mathcal{J}_{\text{Bayes}} = \mathbb{E}_{s_0,\pi_\theta}\Bigg[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \mathbb{E}_{\mathcal{M}\sim p(\mathcal{M}|h_t)}\left[Q^{\pi_\theta}_{\mathcal{M}}(s_t,a_t)\right]\Bigg], \tag{3.1}$$

where we write the history h_t as s_t in π and Q since in LLM reasoning, s_t already encodes the full sequence of prior states and actions, and the reward is unavailable at test time so it's absorbed into θ as a function of s_t. Here, we use the state-action value instead of the advantage since the latter requires multiple Monte Carlo rollouts at each step. By applying the Bayes rule, the posterior satisfies:

$$p(\mathcal{M} \mid h_t) \propto p(\mathcal{M} \mid s_{0:t}) \cdot p(r_{0:t-1} \mid s_{0:t}, a_{0:t-1}, \mathcal{M}),$$

where $p(\mathcal{M}|s_{0:t})$ conditions only on the CoT, excluding rewards, and can be interpreted as the model's probability of outputting solution $y^{\mathcal{M}}_{s_0}$. Interestingly, the second term $p(r_{0:t-1}|s_{0:t}, a_{0:t-1}, \mathcal{M})$ measures the likelihood of observing the rewards $r_{0:t-1}$ under $\mathcal{M}$ given the trajectory $s_{0:t}$. That is, we may write $p(r_t|s_t, a_t, \mathcal{M}) \propto \exp(-\beta|r_t - r_{\mathcal{M}}(s_t, a_t)|)$ with a hyperparameter β to obtain

$$p(r_{0:t-1} \mid s_{0:t}, a_{0:t-1}, \mathcal{M}) \propto \prod_{t'=0}^{t-1} \exp\big(-\beta|r_{t'} - r_{\mathcal{M}}(s_{t'}, a_{t'})|\big),$$

where the proportionality holds since $p(r_{t'}|r_{0:t'-1}, s_{0:t}, a_{0:t-1}, \mathcal{M}) = p(r_{t'}|s_{t'}, a_{t'}, \mathcal{M})$.

Besides, the posterior-weighted value in (3.1) satisfies

$$\begin{aligned}\mathbb{E}_{\mathcal{M}\sim}\left[Q^{\pi_\theta}_{\mathcal{M}}(s_t,a_t)\right] &= \mathbb{E}_{\mathcal{M}\sim q(\mathcal{M}|s_0)}\left[Q^{\pi_\theta}_{\mathcal{M}}(s_t,a_t)\frac{p(\mathcal{M} \mid h_t)}{q(\mathcal{M} \mid s_0)}\right] \\ &= \sum_{i=0}^{|\mathcal{M}|} Q^{\pi_\theta}_{\mathcal{M}_i}(s_t,a_t)p(\mathcal{M}_i \mid h_t),\end{aligned}$$

where the proposal $q(\mathcal{M}|s_0)$ is the uniform distribution over the support of plausible MDPs, defined w.r.t. the ground-truth answer and candidate answers extracted from the model's CoTs. We draw $|\mathcal{M}|$ CoT rollouts of π_θ on prompt s_0 to form $\{\mathcal{M}_i\}_{i=1}^{|\mathcal{M}|}$. The importance ratios are then self-normalized over $\{\mathcal{M}_i\}_{i=1}^{|\mathcal{M}|}$ and the ground-truth MDP $\mathcal{M}_0$ defined w.r.t. $y^*_{s_0}$. By letting $p(\mathcal{M}_i|s_{0:t})$ be the model's state-conditional belief, we obtain:

$$\mathbb{E}_{\mathcal{M}}\left[Q^{\pi_\theta}_{\mathcal{M}}(s_t,a_t)\right] = \sum_{i=0}^{|\mathcal{M}|} \underbrace{Q^{\pi_\theta}_{\mathcal{M}_i}(s_t,a_t)}_{\text{value in } \mathcal{M}_i} \underbrace{\pi_\theta(y^{\mathcal{M}_i}_{s_0}|s_t + \text{¡/think¿})}_{\text{LLM's belief of } \mathcal{M}_i\text{'s plausibility}} \cdot \prod_{t'=0}^{t-1} \underbrace{\exp\big(-\beta|r_{t'} - r_{\mathcal{M}_i}(s_{t'}, a_{t'})|\big)}_{\text{discrepancy b/w obs. \& } \mathcal{M}_i\text{'s prediction}}, \tag{3.2}$$

where $r_{t'}$ is the actual observed reward as defined in (B.1), and $r_{\mathcal{M}_i}$, $Q^{\pi_\theta}_{\mathcal{M}_i}$ are defined in (B.3) w.r.t. the hypothesis $\mathcal{M}_i$. For clarity, we have omitted the normalization constant when computing $p(\mathcal{M}_i|h_t)$ from the last two terms.

BARL offers a principled framework for stitching together plausible strategies, analogous to linearizing best-of-N reasoning, but with guidance on *when* and *how* LLMs should reflectively explore.

Remark 3.1. BARL maximizes the weighted sum of values defined over each hypothesis MDP $\mathcal{M}_i$. The first weighting term $\pi_\theta(y^{\mathcal{M}_i}_{s_0}|\cdot)$ captures LLM's state-conditional belief in the plausibility of $\mathcal{M}_i$. The second product weighting term accumulates the discrepancy between predicted rewards $r_{\mathcal{M}_i}(s_{t'}, a_{t'})$ and observed rewards $r_{t'}$, which serves as a reflective signal for strategy switching by downweighting hypotheses that have high belief probabilities but are unlikely to be optimal.

4 How Bayes-Adaptive RL Helps Generalization: A Didactic Example

In this section, we present a didactic example to show how BARL facilitates test-time generalization. Consider the action space $\mathcal{A}$ that consists of three tokens, $\{0, 1, 2\}$, with one token generated at each timestep. The state is simply $s_{t+1} = a_t$. The objective is to repeat the prompt token three times consecutively within 29 timesteps ($3^3 + 2$ is the

Model	GSM8K	MATH	CollegeMath	OlympiadBench	Average
Qwen2.5-Math-1.5B	40.0	34.1	6.6	21.8	25.6
GRPO	83.9(±0.5)	71.5(±0.4)	45.1(±0.3)	33.4(±0.4)	58.4(±0.3)
Progress	84.8(±0.6)	72.3(±0.4)	45.9(±0.3)	35.6(±0.2)	59.6(±0.2)
BARL	**86.0**(±0.6)	**72.7**(±0.3)	**46.8**(±0.2)	**35.8**(±0.4)	**60.3**(±0.1)
Qwen2.5-Math-7B	59.1	53.7	21.9	19.0	38.4
GRPO	90.3(±0.1)	77.6(±0.4)	47.0(±0.3)	39.0(±0.2)	63.5(±0.2)
Progress	91.1(±0.3)	78.8(±0.1)	47.2(±0.3)	41.1(±0.2)	64.5(±0.2)
BARL	**91.7**(±0.2)	**79.2**(±0.3)	**47.5**(±0.2)	**42.0**(±0.4)	**65.1**(±0.3)

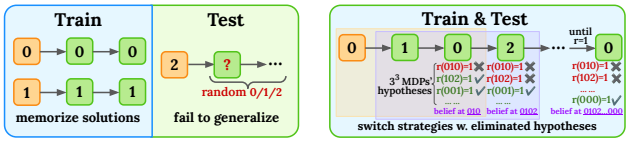
minimal length of a sequence to include all unique triplets). The prompt token is 0 or 1 at training time, and 2 at test time. Episodes terminate when receiving a 1 reward.

This synthetic task mirrors LLM reasoning, where the goal is not only to learn specific strategies (here, generating particular triplets) but also to acquire general problem-solving abilities, such as when and how to switch to new strategies. These capabilities are essential for handling distribution shifts between training and evaluation, a common challenge when developing effective reasoning models.

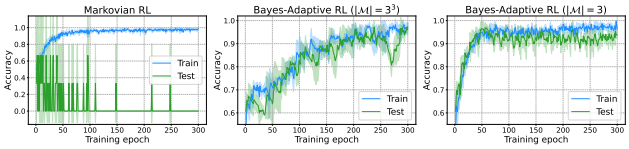
We use a 2-head transformer encoder followed by a linear layer as the policy and train it using the policy gradient from Markovian RL (B.2) and from BARL (3.1). For Markovian RL, the value Q^{π_θ} in the policy gradient is 1 only when $s_{0:T-1}$ contains the rewarding triplet $\text{argmax}_{\text{tri}}\, r(\text{tri})$ of the ground-truth MDP, such as 000 or 111 during training. For BARL, we set $\beta = \infty$ so that $\prod_{t'=0}^{t-1} \exp(-\beta |r_{t'} - r_{\mathcal{M}_i}(s_{t'}, a_{t'})|) = \mathbb{1}(\text{argmax}_{\text{tri}}\, r_{\mathcal{M}_i}(\text{tri}) \notin s_{0:t})$, i.e., the product is 0 when the rewarding triplet of $\mathcal{M}_i$ already appears in $s_{0:t}$ and thus is invalidated, and 1 otherwise. We let the policy's state-conditional belief be $p(\mathcal{M}_i|s_{0:t}) = 1$ for $\mathcal{M}_i$ whose rewarding triplet aligns with the sampled policy action a_t, i.e., $\text{argmax}_{\text{tri}}\, r_{\mathcal{M}_i}(\text{tri}) = s_{t-1:t+1}$, where $s_{t+1} = a_t \sim \pi_\theta(\cdot|s_t)$. Then the posterior-weighted value is

$$\mathbb{E}\big[Q^{\pi_\theta}_{\mathcal{M}}(s_t, a_t)\big] = \mathbb{1}\big(s_{t-1:t+1} \in \{\mathcal{M}_i\}_{i=1}^{|\mathcal{M}|}, s_{t-1:t+1} \notin s_{0:t}\big),$$

where $a_t \sim \pi_\theta(\cdot|s_t)$. The above formulation incentivizes the policy to eliminate hypotheses and switch to new strategies (i.e., new triplets) when the current strategy has been invalidated by earlier attempts up to step t. The difference with (3.2) arises because the agent here is aware of the zero reward associated with unterminated episodes.

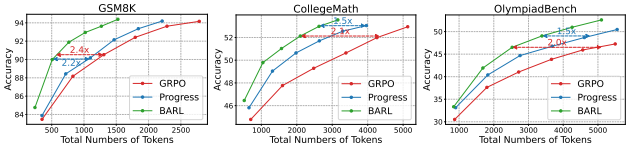


We report the results in the following figure, where accuracies are averaged over 50 completions and the shadow regions are the standard deviation across 3 independent model training runs. The results show that Markovian RL quickly finds and memorizes the training solutions but fails to generalize at test time. In contrast, Bayes-Adaptive RL increases both training and testing accuracies. Furthermore, its accuracy and convergence rate improve when given prior knowledge that rewarding triplets are repeated patterns, i.e., $|\mathcal{M}| = 3$ with $r_{\mathcal{M}_1}(000) = r_{\mathcal{M}_2}(111) = r_{\mathcal{M}_3}(222) = 1$ and all other rewards are zero. This highlights the advantage of more informative candidate sets, underscoring the importance of balancing the *diversity* and *plausibility* of the candidates. Specifically, they should be diverse enough to capture test-time uncertainty, yet constrained to only the most plausible candidates to shrink the hypothesis space.



5 Experiments

We report the pass@1 accuracies in the above table. All models are trained using three random seeds, and we calculate the mean and standard deviation of the resulting accuracies. It can be observed that BARL achieves higher accuracies across most benchmarks and models. It consistently outperforms the two Markovian RL baselines in terms of average accuracy, with the most significant gains observed on challenging benchmarks that demand effective exploration, such as CollegeMath and OlympiadBench. We also evaluate the token efficiency of BARL and baseline models by measuring the total number of tokens required to solve a problem with pass@k in the following figure. We find that BARL achieves higher accuracies with substantially fewer tokens, requiring up to **1.63x** fewer average tokens than the progress baseline, **2x** fewer than GRPO, and over **10x** fewer than the base model. Please refer to Appendix D for more experiment details and ablation studies.



References

Aksitov, R., Miryoosefi, S., Li, Z., Li, D., Babayan, S., Kopparapu, K., Fisher, Z., Guo, R., Prakash, S., Srinivasan, P., et al. Rest meets react: Self-improvement for multi-step reasoning llm agent. *arXiv preprint arXiv:2312.10003*, 2023.

Albalak, A., Phung, D., Lile, N., Rafailov, R., Gandhi, K., Castricato, L., Singh, A., Blagden, C., Xiang, V., Mahan, D., et al. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*, 2025.

Arumugam, D. and Singh, S. Reducing the information horizon of bayes-adaptive markov decision processes via epistemic state abstraction.

Bellman, R. and Kalaba, R. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9, 1959.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Chen, J., Chen, W., and Schneider, J. Bayes adaptive monte carlo tree search for offline model-based reinforcement learning. *arXiv preprint arXiv:2410.11234*, 2024.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Duff, M. O. Monte-carlo algorithms for the improvement of finite-state stochastic controllers: Application to bayes-adaptive markov decision processes. In *International Workshop on Artificial Intelligence and Statistics*, pp. 93–97. PMLR, 2001.

Duff, M. O. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

Gandhi, K., Lee, D., Grand, G., Liu, M., Cheng, W., Sharma, A., and Goodman, N. D. Stream of search (sos): Learning to search in language. *arXiv preprint arXiv:2404.03683*, 2024.

Ghavamzadeh, M., Mannor, S., Pineau, J., Tamar, A., et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

Ghosh, D., Rahme, J., Kumar, A., Zhang, A., Adams, R. P., and Levine, S. Why generalization in rl is difficult: Epistemic pomdps and implicit partial observability. *Advances in neural information processing systems*, 34: 25502–25515, 2021.

Ghosh, D., Ajay, A., Agrawal, P., and Levine, S. Offline rl policies should be trained to be adaptive. In *International Conference on Machine Learning*, pp. 7513–7530. PMLR, 2022.

Guez, A., Silver, D., and Dayan, P. Efficient bayes-adaptive reinforcement learning using sample-based search. *Advances in neural information processing systems*, 25, 2012.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Havrilla, A., Du, Y., Raparthy, S. C., Nalmpantis, C., Dwivedi-Yu, J., Zhuravinskyi, M., Hambro, E., Sukhbaatar, S., and Raileanu, R. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Kazemnejad, A., Aghajohari, M., Portelance, E., Sordoni, A., Reddy, S., Courville, A., and Roux, N. L. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. *arXiv preprint arXiv:2410.01679*, 2024.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Lehnert, L., Sukhbaatar, S., Su, D., Zheng, Q., Mcvay, P., Rabbat, M., and Tian, Y. Beyond a*: Better planning with transformers via search dynamics bootstrapping. *arXiv preprint arXiv:2402.14083*, 2024.

Lidayan, A., Dennis, M. D., and Russell, S. Bamdp shaping: a unified framework for intrinsic motivation and reward shaping. In *The Thirteenth International Conference on Learning Representations*.

Lidayan, A., Dennis, M., and Russell, S. Bamdp shaping: a unified theoretical framework for intrinsic motivation and reward shaping. *arXiv preprint arXiv:2409.05358*, 2024.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Liu, Z., Hu, H., Zhang, S., Guo, H., Ke, S., Liu, B., and Wang, Z. Reason for future, act for now: A principled framework for autonomous llm agents with provable sample efficiency. *arXiv preprint arXiv:2309.17382*, 2023.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Luo, L., Liu, Y., Liu, R., Phatale, S., Lara, H., Li, Y., Shu, L., Zhu, Y., Meng, L., Sun, J., et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2, 2024.

Martin, J. J. *Some Bayesian decision problems in a Markov chain.* PhD thesis, Massachusetts Institute of Technology, 1965.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Qiu, L., Jiang, L., Lu, X., Sclar, M., Pyatkin, V., Bhagavatula, C., Wang, B., Kim, Y., Choi, Y., Dziri, N., et al. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*, 2023.

Qiu, L., Sha, F., Allen, K., Kim, Y., Linzen, T., and van Steenkiste, S. Bayesian teaching enables probabilistic reasoning in large language models. *arXiv preprint arXiv:2503.17523*, 2025.

Qu, Y., Yang, M. Y., Setlur, A., Tunstall, L., Beeching, E. E., Salakhutdinov, R., and Kumar, A. Optimizing test-time compute via meta reinforcement fine-tuning. *arXiv preprint arXiv:2503.07572*, 2025.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Setlur, A., Nagpal, C., Fisch, A., Geng, X., Eisenstein, J., Agarwal, R., Agarwal, A., Berant, J., and Kumar, A. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*, 2024.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P. J., Harrison, J., Lee, J., Xu, K., et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Tang, Z., Zhang, X., Wang, B., and Wei, F. Mathscale: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*, 2024.

Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Wang, H., Hao, S., Dong, H., Zhang, S., Bao, Y., Yang, Z., and Wu, Y. Offline reinforcement learning for llm multi-step reasoning. *arXiv preprint arXiv:2412.16145*, 2024a.

Wang, P., Li, L., Shao, Z., Xu, R., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023a.

Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., and Goodman, N. D. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023b.

Wang, Z., Li, Y., Wu, Y., Luo, L., Hou, L., Yu, H., and Shang, J. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. *arXiv preprint arXiv:2402.02658*, 2024b.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wei, Y., Duchenne, O., Copet, J., Carbonneaux, Q., Zhang, L., Fried, D., Synnaeve, G., Singh, R., and Wang, S. I. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. *arXiv preprint arXiv:2502.18449*, 2025.

Xiang, V., Snell, C., Gandhi, K., Albalak, A., Singh, A., Blagden, C., Phung, D., Rafailov, R., Lile, N., Mahan, D., et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. *arXiv preprint arXiv:2501.04682*, 2025.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., and Zhang, Z. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Yeo, E., Tong, Y., Niu, M., Neubig, G., and Yue, X. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Fan, T., Liu, G., Liu, L., Liu, X., et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yuan, Z., Yuan, H., Li, C., Dong, G., Lu, K., Tan, C., Zhou, C., and Zhou, J. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Zelikman, E., Harik, G., Shao, Y., Jayasiri, V., Haber, N., and Goodman, N. D. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

Zeng, W., Huang, Y., Liu, Q., Liu, W., He, K., Ma, Z., and He, J. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

Zhang, S., Yu, D., Sharma, H., Zhong, H., Liu, Z., Yang, Z., Wang, S., Hassan, H., and Wang, Z. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024.

Zhong, H., Yin, Y., Zhang, S., Xu, X., Liu, Y., Zuo, Y., Liu, Z., Liu, B., Zheng, S., Guo, H., et al. Brite: Bootstrapping reinforced thinking process to enhance language model reasoning. *arXiv preprint arXiv:2501.18858*, 2025.

A Related Work

LLM Reasoning. As an emerging capability of model scale, LLMs can generate intermediate CoTs to solve complex reasoning tasks (Wei et al., 2022; Kojima et al., 2022) and scale test-time performance by allocating more thinking tokens (Snell et al., 2024; Brown et al., 2024). Early efforts enhanced LLM reasoning via supervised fine-tuning on human-annotated data (Cobbe et al., 2021; Yue et al., 2023; Yu et al., 2023) or linearized search traces (Lehnert et al., 2024; Gandhi et al., 2024). However, due to the distribution shift between LLM responses and curated data, LLM-generated data has proven effective through rejection sampling (Dong et al., 2023; Yuan et al., 2023) by filtering out low-quality rationales (Zelikman et al., 2022; 2024) or with EM iterations (Singh et al., 2023; Zhong et al., 2025). Recently, RL has gained increasing interest for improving reasoning (Aksitov et al., 2023; Havrilla et al., 2024; Wang et al., 2024a; Shao et al., 2024). Process rewards (Uesato et al., 2022; Lightman et al., 2023) with Monte Carlo unrolls (Kazemnejad et al., 2024; Wang et al., 2023a; 2024b; Luo et al., 2024) offer finer-grained feedback but are computationally expensive. Outcome-reward RL (Guo et al., 2025) demonstrates emergent deliberative reasoning abilities such as self-reflection. Yet, limited work has investigated the underlying mechanisms of such behaviors. In fact, recent findings suggest that reflections do not consistently emerge from RL training and exhibit weak correlation with performance (Liu et al., 2025). Similar to our work, (Xiang et al., 2025; Qu et al., 2025) also study the generalization of LLMs, from a meta-RL (Duan et al., 2016) perspective: (Xiang et al., 2025) justify deliberative reasoning as providing extra contexts, and (Qu et al., 2025) use progress reward (Setlur et al., 2024) to reduce regret in outcome-reward RL. Our method differs from (Xiang et al., 2025) in that we ground reflective reasoning in environment rewards, rather than relying solely on the internal CoT states generated by the model itself. Compared to (Qu et al., 2025), which rewards golden strategies that make progress towards the correct answer, BARL additionally encourages exploring plausible strategies under the Bayesian framework, allowing it to account for uncertainty during both training and testing. We experimentally compare with a variant of (Qu et al., 2025) that estimates progress using answer probability differences. Besides, unlike (Wang et al., 2023b; Qiu et al., 2023) that manually design hypothesis proposal–selection pipelines, our method achieves this through more principled RL optimization.

Reinforcement Learning. Conventional RL explores only during training, e.g. via ϵ-greedy noise, and exploits the optimal deterministic policy when deployed. Exceptions include works that explicitly optimize maximum entropy objectives (Haarnoja et al., 2018) to learn stochastic policies, primarily to accelerate *training* convergence in settings where evaluation remains in-distribution, such as robotic control. Bayes-Adaptive RL (Bellman & Kalaba, 1959; Duff, 2002; Guez et al., 2012; Ghavamzadeh et al., 2015; Lidayan et al., 2024; Zhang et al., 2024; Liu et al., 2023) has been studied to pursue the optimal exploration-exploitation trade-off in uncertain environments to improve generalizability. When the true MDP identity is latent and must be inferred from interaction (states, actions, rewards), Bayesian RL connects naturally to Partially Observable MDPs (Duff, 2001; Ghosh et al., 2021). Exact solutions of the Bayesian RL objective are often intractable, prompting the development of approximate methods (Guez et al., 2012; Arumugam & Singh; Chen et al., 2024). In our work, we adopt policy gradient that operates over candidate answers, which differs from (Ghosh et al., 2022) that leverage value ensembles in offline RL and (Qiu et al., 2025) that applies SFT on an oracle Bayesian model's outputs.

B Problem Formulation

LLM Reasoning via RL. To enable the LLM policy π_θ to reason, we first consider the finite-horizon MDP defined by the state space $\mathcal{S}$, action space $\mathcal{A}$, horizon T, and reward function $r(s, a)$, where $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Here, the initial state s_0 is the prompt, and the action a_t is the t-th step of the CoT, which can be either separated by special tokens (Wang et al., 2023a) or defined as a fixed length of reasoning tokens (Luo et al., 2024). We adopt the latter definition due to its simplicity. The state transition is deterministic by appending the new reasoning step, i.e., $s_{t+1} = s_t + a_t$. Prior work (Uesato et al., 2022; Guo et al., 2025) employs an outcome-level reward verifier$(s_T, y^*_{s_0})$, which uses a verifier to perform a regular expression match (either 0 or 1) between s_T and the ground-truth answer $y^*_{s_0}$ corresponding to the prompt s_0. We extend this sparse-reward setting by incorporating a *progress* reward (Setlur et al., 2024; Qu et al., 2025), which quantifies the increase of the model's probability of outputting $y^*_{s_0}$ after appending a at the CoT s, i.e., for $0 \le t \le T-1$:

$$r(s_t, a_t) = \pi_\theta(y^*_{s_0} \mid s_t + a_t + \text{¡/think¿}) - \pi_\theta(y^*_{s_0} \mid s_t + \text{¡/think¿}), \tag{B.1}$$

where ¡/think¿ is the end sign of thinking, such as the answer elicitation prompt "Based on the above reasoning, the answer is `\boxed`" that we adopt. Compared to Monte-Carlo process rewards (Luo et al., 2024; Qu et al., 2025), (B.1) is computationally efficient by avoiding multiple branched rollouts at each step, and the KV cache of $s_{0:T}$ from CoT generations can also be reused.

For the Markovian RL objective $\mathcal{J}_{\text{RL}}(\pi_\theta) := \mathbb{E}_{s_0,\pi_\theta}[\sum_{t=0}^{T-1} r(s_t, a_t) + \text{verifier}(s_T, y^*_{s_0})]$, this reward definition allows us to

use telescoping in a way similar to reward shaping (Ng et al., 1999) to obtain

$$\operatorname*{argmax}_{\pi_\theta} \mathcal{J}_{\mathrm{RL}}(\pi_\theta) = \operatorname*{argmax}_{\pi_\theta} \mathbb{E}_{s_0,\pi_\theta}\left[\pi_\theta(y^*_{s_0} \mid s_0 + a_{0:T-1} + \text{¡/think¿}) + \text{verifier}(s_{T+1}, y^*_{s_0})\right],$$

i.e., the optimal Markovian policy generates $a_{0:T-1}$ to maximize the likelihood of the ground-truth $y^*_{s_0}$ and its verifier-evaluated correctness. The gradients for Markovian policies are

$$\nabla_\theta \mathcal{J}_{\mathrm{RL}}(\pi_\theta) = \mathbb{E}_{s_0,\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot Q^{\pi_\theta}(s_t, a_t)\right], \tag{B.2}$$

where Q^{π_θ} is the state-action value or advantage function (Schulman et al., 2015). The above setups consider the case when the environment is predefined with certainty. The definitions naturally extend to any MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, r_{\mathcal{M}}, T)$ where $r_{\mathcal{M}}$ is defined w.r.t. the answer $y^{\mathcal{M}}_{s_0}$. The Q-value is then

$$\begin{aligned} Q^{\pi_\theta}_{\mathcal{M}}(s_t, a_t) &= \mathbb{E}_{\pi_\theta}\left[\sum_{t'=t}^{T-1} r_{\mathcal{M}}(s_{t'}, a_{t'}) + \text{verifier}(s_T, y^*_{s_0})\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\pi_\theta(y^{\mathcal{M}}_{s_0} \mid s_t + a_{t:T-1} + \text{¡/think¿}) - \pi_\theta(y^{\mathcal{M}}_{s_0} \mid s_t + \text{¡/think¿}) + \text{verifier}(s_T, y^*_{s_0})\right]. \end{aligned} \tag{B.3}$$

Reflective Exploration. We define reflective exploration as the pattern in which the LLM backtracks to a prior state after an exploratory step to take different actions at that state. Specifically, a natural language reflective reasoning step such as "Let's reconsider the geometric relationship" corresponds to a backtracking action that semantically disregards the previous one or more steps. We illustrate this using a binary search tree as in the right figure: for the trajectory $s_0 s_1 s_2 s_1 s_3$, $s_1 s_2$ is an exploration step, $s_2 s_1$ is a reflective step that signals the strategy switch from s_2 to s_3, and the "geometric relationship" in the above example originates from s_1.

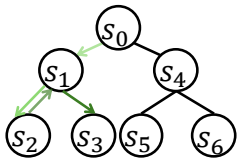


Figure 1. An example of reflective reasoning.

C Proof of Theorem 2.2

Proof. Consider a full binary tree of depth T^*, with leaf set $\mathcal{L}$ of size $|\mathcal{L}| = 2^{T^*}$. The states are the tree nodes, with the initial state fixed as the root node. The actions include the moves from the parent nodes to the child nodes as well as a resetting action from the leaf node to the root node. The rewards are not known and differ in different MDP hypotheses. Specifically, the reward is defined as $r_{\mathcal{M}_i}(s) = \mathbb{1}(s = s_i)$, where s_i is a unique leaf node for $\mathcal{M}_i$. The prior of MDPs is $p(\mathcal{M}_i) = 1/|\mathcal{L}|$, where $i = 1, \cdots, |\mathcal{L}|$. The cumulative return is undiscounted with the minimum traverse steps as the horizon T. The episode terminates once the agent receives a 1 reward.

For any Markovian policy π, define $f(s) = p(\exists t \geq 0 : s_t = s \mid \pi)$. We define p_s as the probability that π goes left at an internal node s. By the Markov property, for any left and right children s_L and s_R of node s, it holds that $f(s_L) = f(s)p_s$ and $f(s_R) = f(s)(1 - p_s)$. Thus, $f(s_L) + f(s_R) = f(s)$. By a simple induction on depth, it follows that at every depth d, $\sum_{s:\text{depth}(s)=d} f(s) = 1$. In particular, $\sum_{l\in\mathcal{L}} f(l) = 1$. Since the total return under $\mathcal{M}_i$ is $V(\pi \mid \mathcal{M}_i) = p(\pi \text{ ends at leaf } l_i) = f(l_i)$, the expected return for the optimal Markovian policy is

$$\frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|} V(\pi \mid \mathcal{M}_i) = \frac{1}{2^{T^*}}\sum_{l\in\mathcal{L}} f(l) = \frac{1}{2^{T^*}}.$$

Consider the following deterministic Bayes-adaptive policy. At the root node, the policy picks any leaf l with positive posterior, i.e., $p(\mathcal{M}_l \mid h_t) > 0$, then follows the unique shortest path to l. Two possible outcomes can occur: if the ground-truth reward $r(l) = 1$, then the episodes terminates with the collected reward; if $r(l) = 0$, then the agent eliminates the hypothesis $\mathcal{M}_l$ from the posterior by setting $p(\mathcal{M}_l \mid h_{t:t+T^*}) = 0$, and returns to the root to repeat the process on the remaining leaves. By construction, the expected return of this Bayes-Adaptive policy is 1, which is an exponential improvement in T^* over the $1/2^{T^*}$ return of the optimal Markovian policy. □

D Experiment Details

D.1 Experiment Setups

In addition to the synthetic experiment in Section 4, we evaluate BARL on LLM math problem-solving tasks. We implement BARL across various models, including Qwen2.5-Math-1.5B, Qwen2.5-Math-7B (Yang et al., 2024), and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025). Training is conducted on the Big-Math dataset (Albalak et al., 2025), and evaluation is performed on four benchmarks: GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), CollegeMath (Tang et al., 2024), and OlympiadBench (He et al., 2024). During training, the maximum prompt length is set to 512 and the maximum response length is set to 1024. We exclude AIME and AMC from evaluation due to their substantially longer context requirements, e.g., DeepSeek-R1-Distill-Llama-8B has average response lengths of 2008 and 1886 tokens on AIME 2024 and AMC 2023, respectively. For BARL, we set $\beta = 1$ and $|\mathcal{M}| = 5$.

We compare BARL against two Markovian RL baselines that span outcome-reward and process-reward RL. For the outcome-reward GRPO baseline, we set its group size to 5 for a fair comparison with BARL and, after performing a grid search over the KL-divergence coefficients $[0, 0.001, 0.005, 0.01]$, adopt 0.005 as it yields the best overall performance across all benchmarks. For the process-reward baseline, we adapt a variant of MRT (Qu et al., 2025) by integrating the progress reward defined in (B.1) into the outcome reward, which we refer to as *progress* in the following sections. For all algorithms, we set the training and rollout batch sizes to 128 and 1024, respectively. We train the Qwen and Llama models for 110 and 60 iterations, respectively, defined w.r.t. the rollout batches. The temperature during online sampling is 1.0 and is 0.0 during evaluation. For both BARL and the progress baseline, we set the number of tokens for each reasoning step as 128.

D.2 Ablation Studies

Reflective Reasoning Behaviors. To qualitatively assess the improved token efficiency of BARL, we analyze the frequency of reflective behaviors across problems of varying difficulty levels, as shown in Figure 2. For each problem, we sample 6 responses per model and define its difficulty level by the number of incorrect responses. We use keyword-based detections (Liu et al., 2025; Yeo et al., 2025) to identify whether self-reflections appear in a response, and a problem is considered to exhibit self-reflection if at least one of its responses is identified. It can be observed that both models display fewer reflections on easier problems, and the base model exhibits a higher frequency of reflections despite achieving lower accuracies. This result reveals the weak correlation between the performance of LLMs and the response length or the frequency of reflections. Rather, the effectiveness of thinking tokens and the efficiency of explorations are the determining factors, which we study in the following ablation.

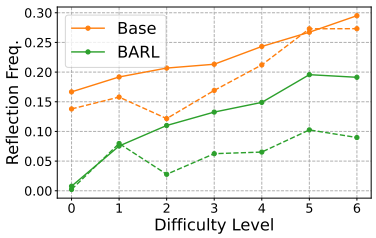


Figure 2. Results on GSM8K (dashed) and MATH (solid).

Effectiveness of CoTs. We measure the effectiveness of the CoTs produced by different models by calculating the average Bayesian state-action values at each timestep, which naturally captures both the exploration and the exploitation aspects of the actions. Specifically, the Bayesian value is defined as $Q^\pi(b_t, s_t, a_t) = \mathbb{E}_{\pi, \mathcal{M}\sim b_t}[r_{\mathcal{M}}(s_t, a_t) + Q^\pi(b_{t+1}, s_{t+1}, a_{t+1})]$, where the belief $b_t = p(\mathcal{M}|h_t)$. Unlike standard Q-values, the Bayesian Q-value not only incorporates the expected returns (exploitation) but also captures the value of information gained through belief updates (exploration).

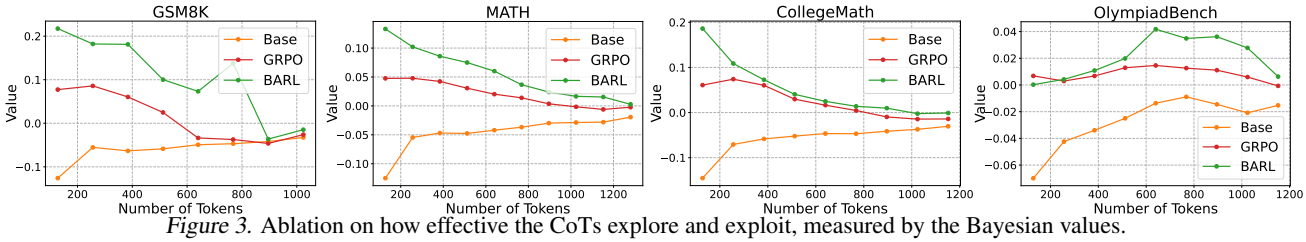


Figure 3. Ablation on how effective the CoTs explore and exploit, measured by the Bayesian values.

The results are reported in Figure 3. We observe that the actions from the BARL model exhibit consistently higher Bayesian values compared to those of the GRPO and base models, indicating more effective exploration and exploitation. On more challenging benchmarks such as OlympiadBench, exploratory gains peak midway through the CoTs after an early phase of uncertainty reduction. Moreover, the result also explains our earlier observations on token efficiency and reflective behaviors. Although the base model exhibits more self-reflections, these are likely superficial or stylistic patterns due to their low exploration efficiency for gathering informative contexts during evaluation.

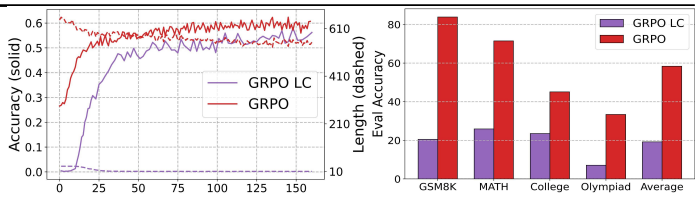


Figure 4. Training accuracies, lengths and eval results.

Markovian RL Optimality. We train a length-controlled (LC) GRPO with a maximum 32 response length over multiple epochs. Figure 4 shows the evolution of the training accuracy and response length. The rapidly decreasing length of GRPO LC indicates that it learns to skip CoT generation and emit only the final answer. Its asymptotic training accuracy matches that of GRPO (max length 1024). This result supports Theorem 2.1: Markovian RL can achieve optimality by merely memorizing solutions without reflective reasoning. Such policies, however, generalize poorly during evaluation.

Key Experiment Findings

BARL consistently outperforms Markovian RL baselines with superior token efficiency. Performance correlates with the effectiveness of reflective explorations, rather than their frequency. Optimality in Markovian RL can be attained by policies that memorize training solutions yet fail to generalize, with no guarantees on the emergence of self-reflections.

D.3 Evaluation Results: Accuracies

In this section, we provide the evaluation accuracy results for Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, and DeepSeek-R1-Distill-Llama-8B in Figure 5, 6, and 7, respectively. In addition to the benchmark scores in Table **??**, we also report the performance of the models on AIME 2024 and AMC 2023. It can be observed that BARL outperforms Markovian RL baselines in terms of both accuracy and convergence rate on most of the reported benchmarks. Again, the performance on AIME and AMC benchmarks may be further enhanced by choosing harder training data with increased response length, especially for R1-Distill-Llama-8B fine-tuned models whose initial average response lengths on AIME and AMC (2008 and 1886, respectively) exceed the maximum training length (1024).

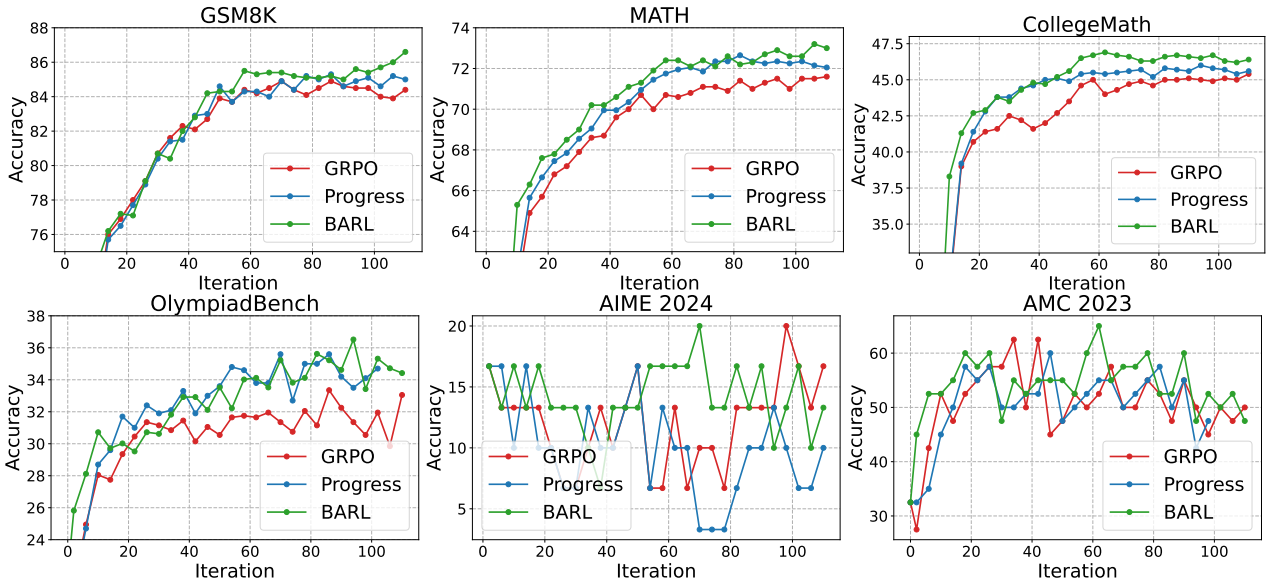


Figure 5. Average evaluation accuracies over training iterations for Qwen2.5-Math-1.5B models.

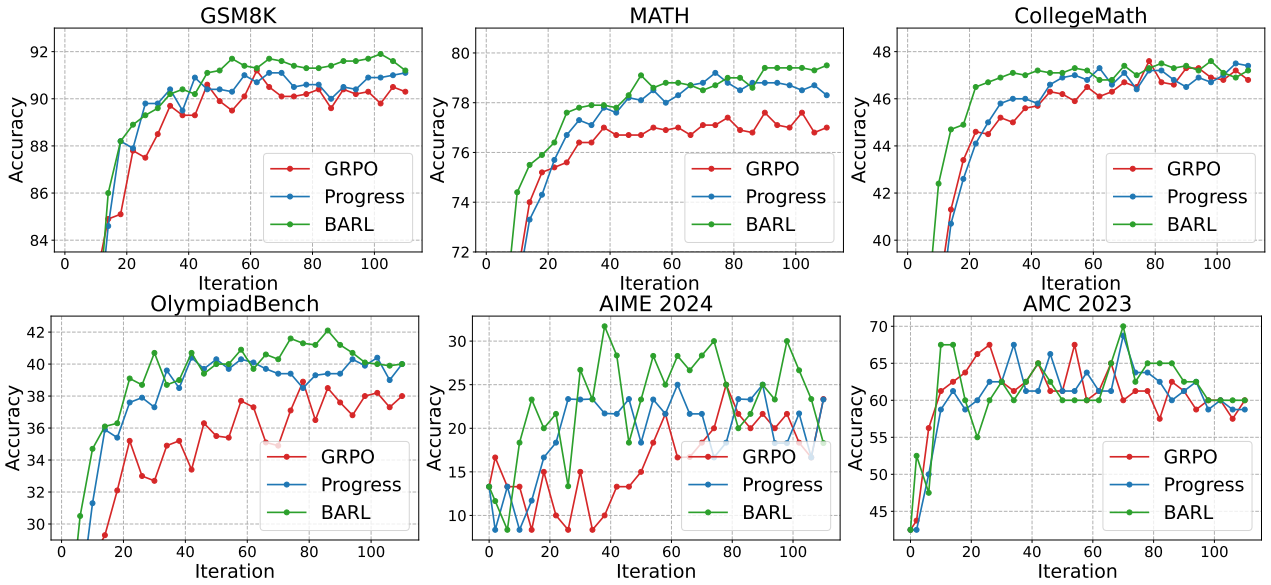


Figure 6. Average evaluation accuracies over training iterations for Qwen2.5-Math-7B models.

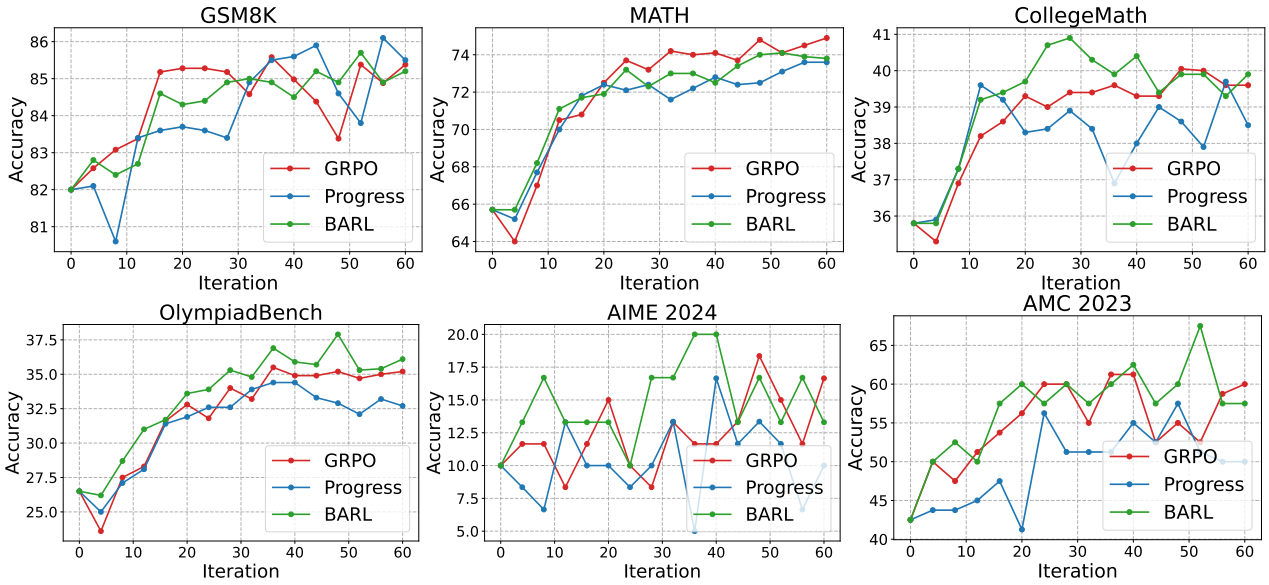


Figure 7. Average evaluation accuracies over training iterations for R1-Distill-Llama-8B models.

D.4 Evaluation Results: Response Lengths

In this section, we present the evolution of evaluation response lengths over training iterations for Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, and DeepSeek-R1-Distill-Llama-8B, shown in Figures 8, 9, and 10, respectively. Across most benchmarks, response lengths tend to decrease as training progresses for all algorithms. The response length during training has a very similar trend to that during evaluation. This trend arises because all three models exhibit reflective behaviors, such as self-evaluation and backtracking, that introduce redundant tokens and lengthen responses. As shown in Figure 3, these behaviors are likely superficial or stylistic patterns with limited effectiveness. An exception is AIME, where some models maintain consistently long responses due to the benchmark's intrinsic requirement for extended reasoning, even under optimal non-reflective policies.

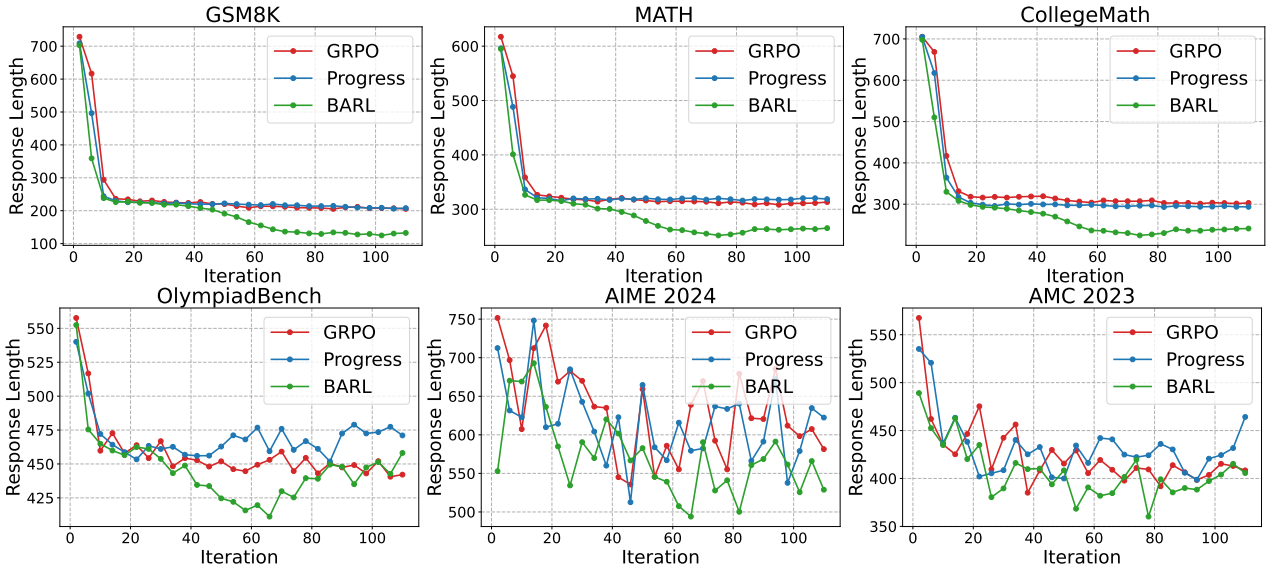


Figure 8. Average evaluation response lengths over training iterations for Qwen2.5-Math-1.5B models.

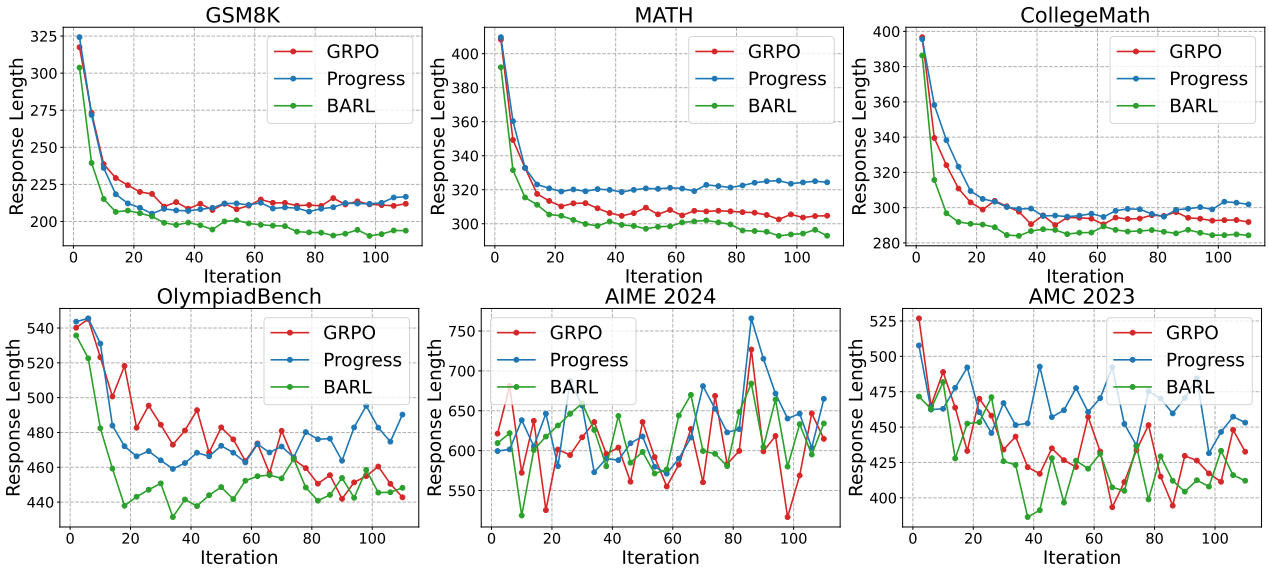


Figure 9. Average evaluation response lengths over training iterations for Qwen2.5-Math-7B models.

Since the models used in the main experiments already exhibit lengthy CoTs with reflective patterns, we further implement BARL on the Llama-3.2-3B-Instruct model, which displays fewer self-reflections. The results are presented in Figure 11. Initially, the response length decreases as the base model tends to produce excessively long reasoning traces, often exceeding ten steps, which are pruned during early training. Subsequently, the response length increases, as a result of plausible strategy stitching.

D.5 Token Efficiency without Greedy Decoding Outputs

In Figure 12, we present the pass@k accuracies from an ablation similar to Section D.2, except that greedy decoding outputs are excluded when computing token counts and accuracies. We observe that the base and GRPO models are less robust under a sampling temperature of 1.0, resulting in significantly lower pass@1 accuracies compared to greedy decoding. This degradation may stem from the fragility of their CoTs, which often exhibit stylistic but unproductive self-reflection and backtracking behaviors.

D.6 Some Unsuccessful Attempts

As a straightforward implementation of the Bayes-Adaptive RL policy gradient in (3.1), we explored using value ensembles to estimate the posterior-weighted value $\mathbb{E}_{\mathcal{M}\sim p(\mathcal{M}|h_t)}[Q^{\pi_\theta}_{\mathcal{M}}(s_t, a_t)]$. Specifically, we trained an ensemble of state-action value functions to capture epistemic uncertainty. We experimented with two approaches to constructing the ensemble: (1) fine-tuning multiple linear value heads on disjoint data subsets, each paired with chain-of-thought (CoT) trajectories and outcome rewards; and (2) applying Bayesian LoRA. However, both methods failed to effectively capture epistemic uncertainty, likely because they fine-tune only a small subset of LLM parameters, which is insufficient to fully represent the uncertainty. While maintaining independent value models may better capture this uncertainty, doing so incurs substantial computational cost. We leave the development of more efficient implementations to future work.

E Conclusion

Large Language Models (LLMs) trained via Reinforcement Learning (RL) have exhibited emergent behaviors such as self-reflective reasoning. Yet in conventional Markovian RL, exploration is confined to the training phase to identify action sequences that maximize cumulative reward, and resorts to pure exploitation at test time. Besides, the Markov assumption indicates the dependency on history only through the state. Thus, Markovian RL neither ensures the emergence of reflective exploration nor explains its benefits during testing. We propose to fill this gap with Bayes-Adaptive RL, which explicitly considers test-time performance by maximizing the expected return under a posterior of MDPs. Within this framework, we propose BARL, a novel algorithm for LLM reasoning that provides principled guidance for when and how to engage in reflective exploration. BARL enables efficient exploration through hypothesis elimination and strategy switching. Our

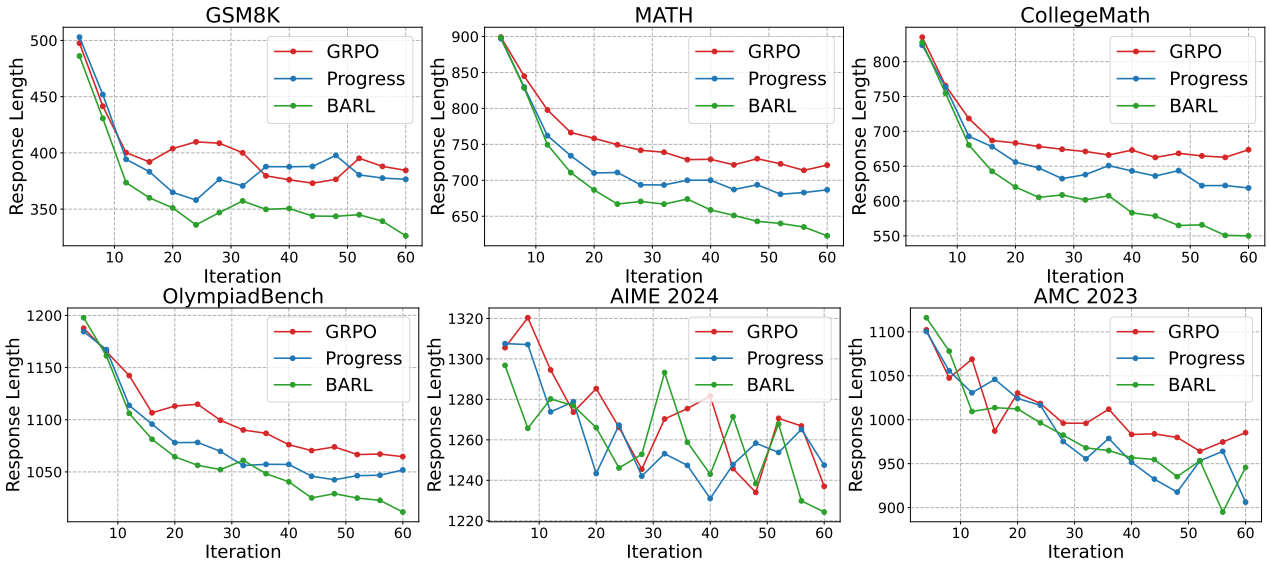
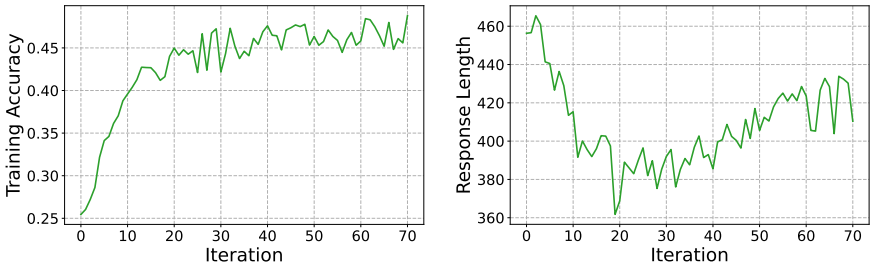


Figure 10. Average evaluation response lengths over training iterations for R1-Distill-Llama-8B models.



	Base	BARL
GSM8K	69.4	**73.8**
MATH	47.2	**50.5**
CollegeMath	30.8	**33.6**
Olympiad	16.4	**19.1**
AIME 24	3.3	**13.3**
AMC 23	27.5	**35.0**

Figure 11. Results of BARL fine-tuned on Llama-3.2-3B-Instruct. **(Left)** Training accuracy and **(Middle)** response length. **(Right)** Evaluation results.

experiments are conducted on both synthetic and mathematical reasoning tasks, where we show that BARL outperforms Markovian RL algorithms at test time and its exploration is more efficient. As future work, we plan to extend our approach to broader domains, such as coding and agentic tasks.

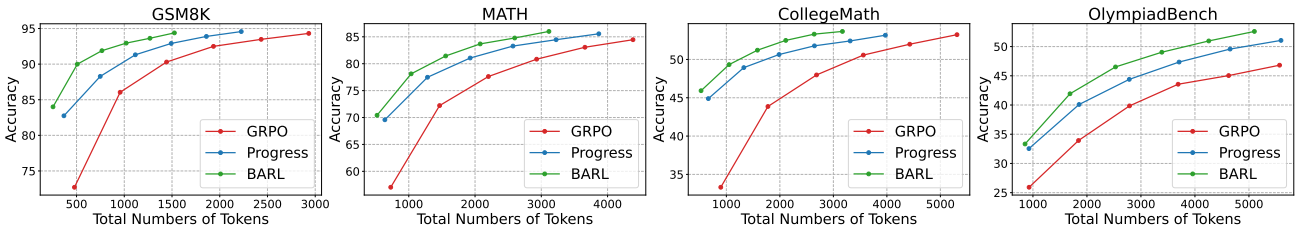


Figure 12. Ablation on token efficiency and pass@k accuracies with sampling temperature= 1. GRPO and the base models are less robust to temperatures.