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

Shenao Zhang[1], Yaqing Wang[2], Yinxiao Liu[2], Tianqi Liu[2],
Peter Grabowski[3], Eugene Ie[3], Zhaoran Wang[†1], Yunxuan Li[†3]

[1]Northwestern University, [2]Google DeepMind, [3]Google

We study ***why***, ***how***, and ***when*** LLMs should self-reflect and explore *at test time*
—questions that conventional Markovian RL cannot fully answer.

Why to self-reflect?

❌ Markovian RL ensures *neither* the emergence of self-reflection …

- Optimal Markovian policies can simply memorize training solutions.

❌ … *nor* the test-time benefits of self-reflection.

- Markovian RL *explores* during training and *exploits* during testing.
- Such trial-and-error exploration lacks incentives to collect extra context and backtrack.

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

✅ Bayes-adaptive RL: info-gathering exploration to reduce the MDP's uncertainty.

How to self-reflect?

We propose BARL. It provides step-level guidance to stitch plausible strategies (by sampling |M| responses for each prompt), akin to linearized BoN.

$$\mathbb{E}_{M\sim p(M|h_t)}\left[Q_M^{\pi_\theta}(s_t,a_t)\right] = \sum_{i=0}^{|M|} \underbrace{Q_{M_i}^{\pi_\theta}(s_t,a_t)}_{\text{value in } M_i} \cdot \underbrace{\pi_\theta(y_{s_0}^{M_i} \mid s_t + \text{</think>})}_{\text{LLM's belief in } M_i\text{'s plausibility}} \cdot \prod_{t'=0}^{t-1} \underbrace{\exp\left(-\beta \left| r_{t'} - r_{M_i}(s_{t'},a_{t'}) \right|\right)}_{\text{consistency b/w obs. \& } M_i\text{'s pred.}}$$

Sum of M_i's Q, weighted by LLM's belief *and* consistency between true MDP & M_i.

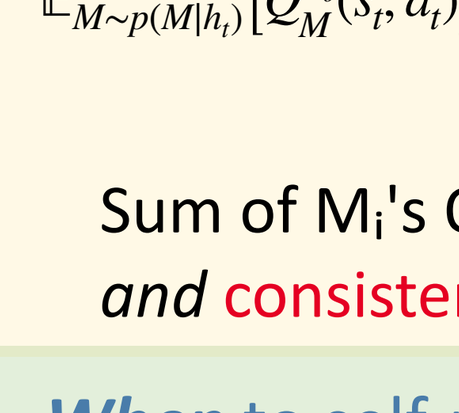
$$\nabla_\theta \mathcal{J}_{\text{bayes}} = \mathbb{E}_{s_0,\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \mathbb{E}_{M\sim p(M|h_t)}\left[Q_M^{\pi_\theta}(s_t,a_t)\right]\right]$$

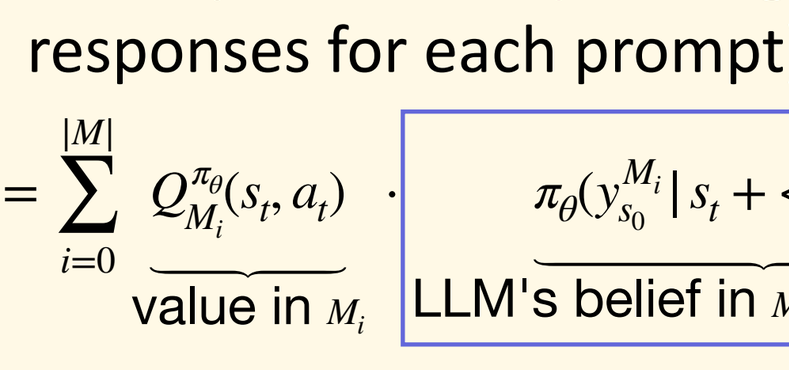
When to self-reflect?

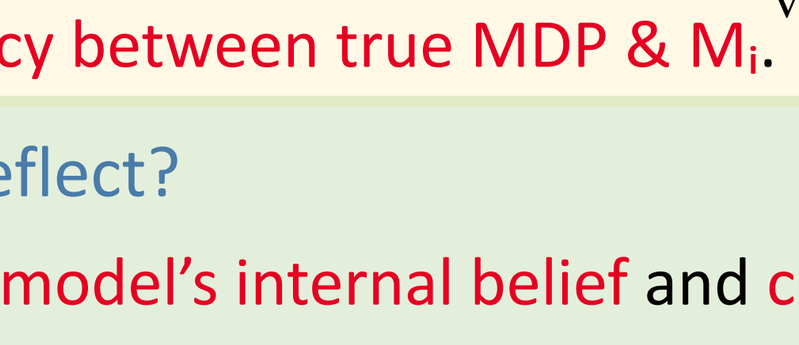
Reflect when model's internal belief and cumulative reward feedback mismatch.

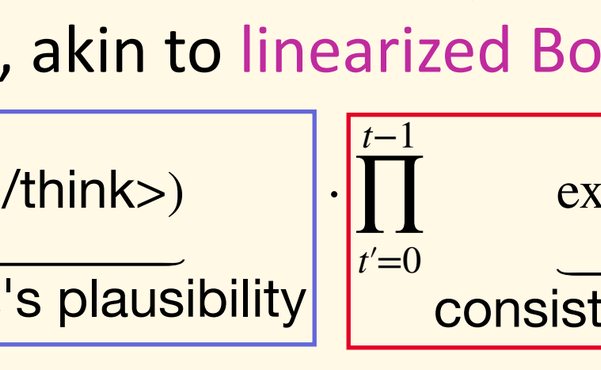
- Time to switch strategy if it is unlikely to be optimal given previous observations!

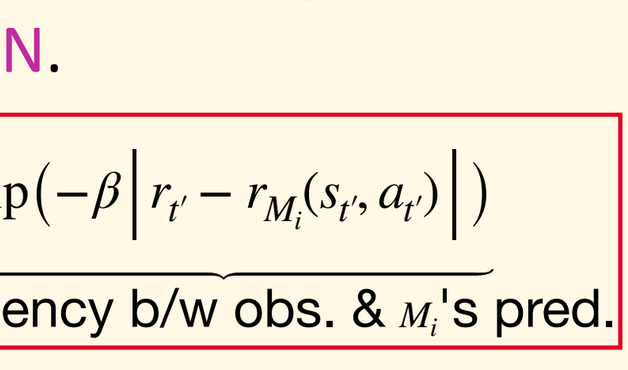
Experiments











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)
R1-Distill-Llama-8B	82.0	65.7	35.8	26.5	52.5
GRPO	85.7(±0.5)	**74.3**(±0.4)	39.6(±0.3)	36.0(±0.5)	58.9(±0.4)
Progress	**85.9**(±0.4)	73.9(±0.4)	39.8(±0.5)	35.3(±0.2)	58.7(±0.4)
BARL	85.4(±0.6)	73.9(±0.6)	**40.4**(±0.3)	**37.2**(±0.5)	**59.3**(±0.4)

arXiv