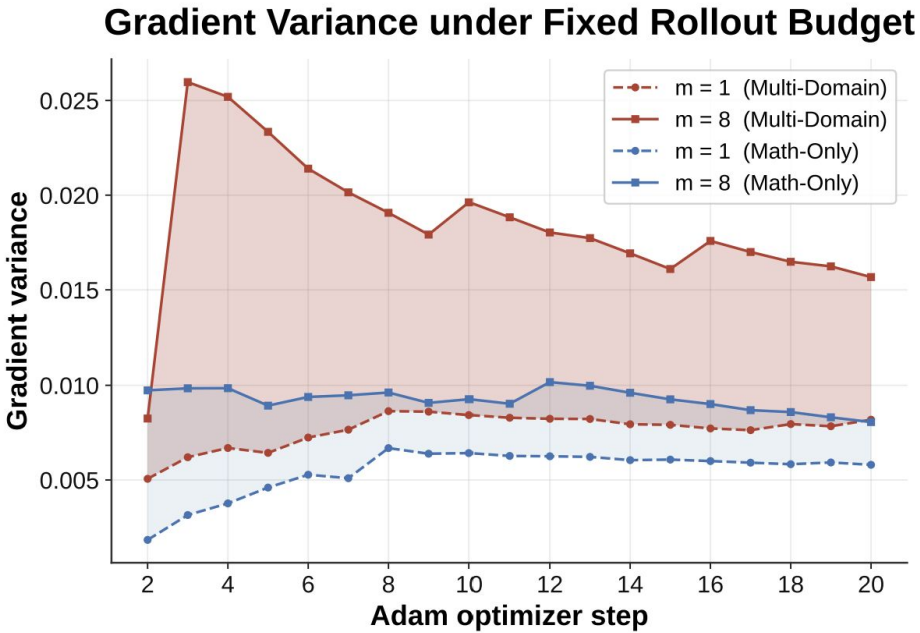
Your Language Model is Its Own Critic: Reinforcement Learning with Value Estimation from Actor's Internal States

Yunho Choi*, Jongwon Lim*, Woojin Ahn, Minjae Oh, Jeonghoon Shim, Yohan Jo
Seoul National University
NeurIPS 2026

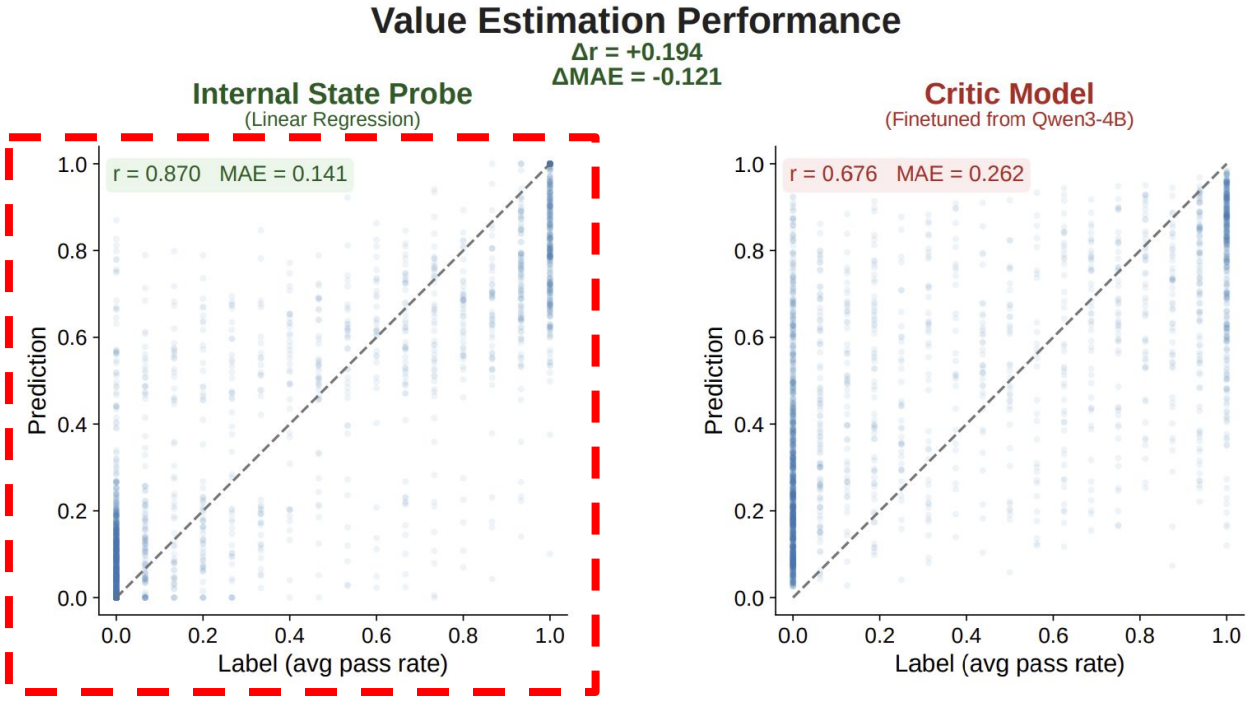
Challenge

GRPO generates multiple responses per prompt,
leaving fewer prompts per batch and potentially higher gradient variance.



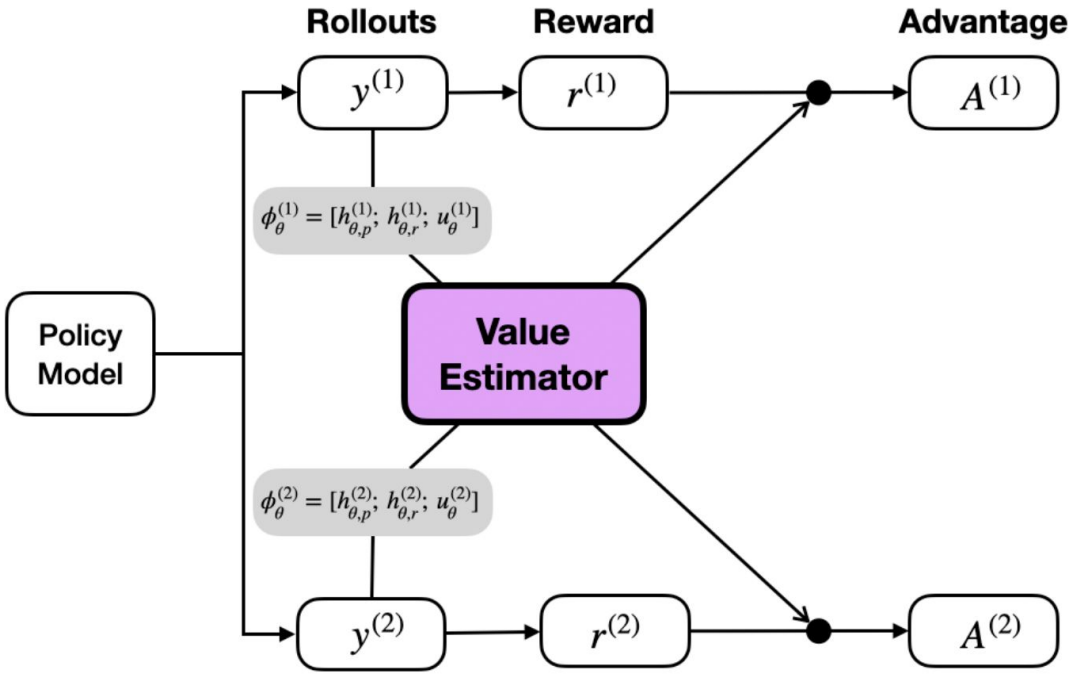
Motivation

A simple linear probe on the actor **estimates values more accurately than a separate critic model**.



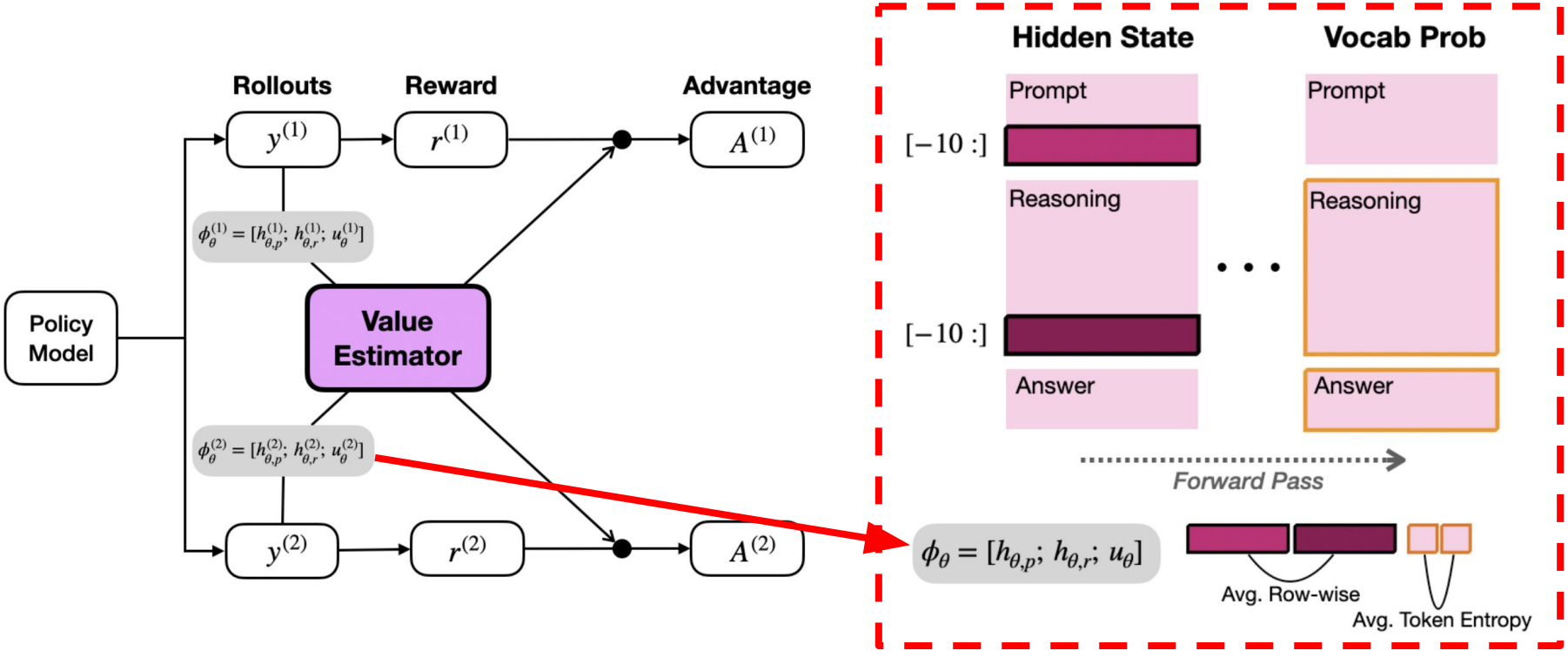
Method

Policy Optimization with Internal State Value Estimation (POISE)



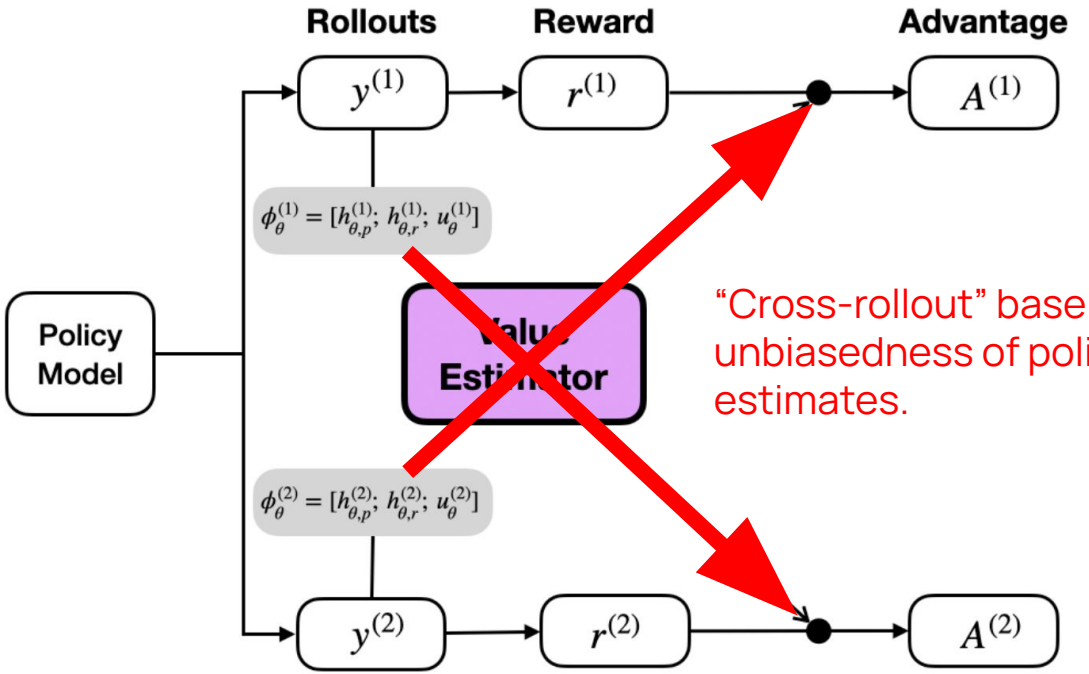
Method

Policy Optimization with Internal State Value Estimation (POISE)



Method

Policy Optimization with Internal State Value Estimation (POISE)



"Cross-rollout" baseline preserves unbiasedness of policy-gradient estimates.

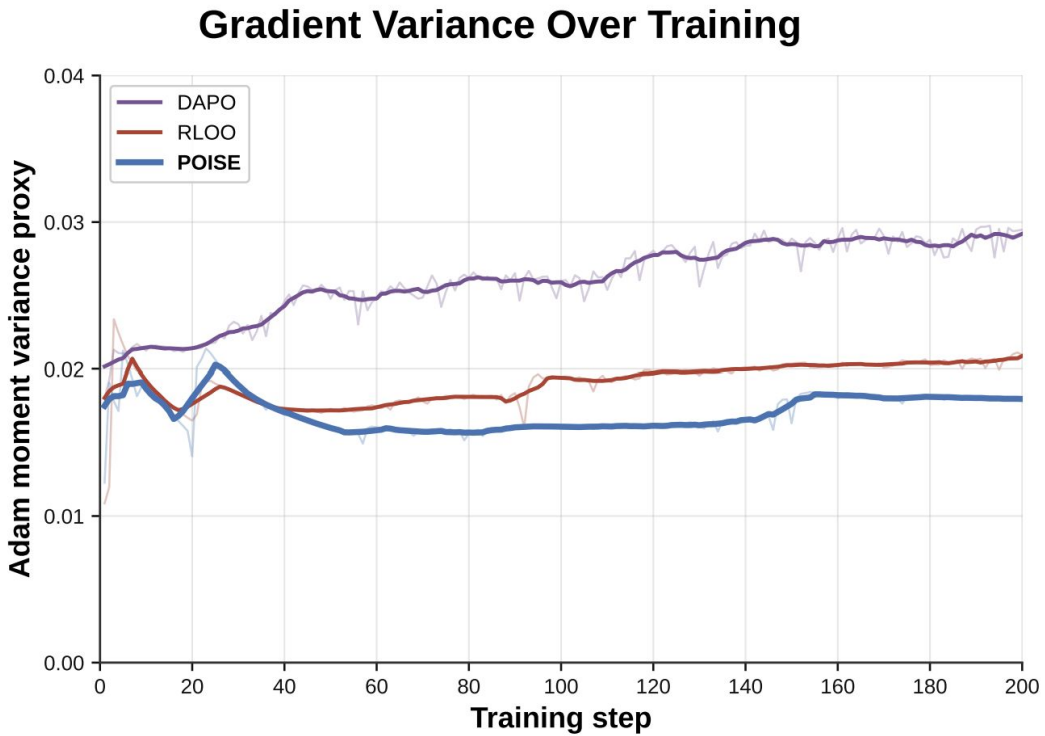
Results

POISE **outperforms other RL baselines** on average across domains.

Domain	Qwen3-4B				OLMo-3-7B-Instruct-DPO			
	Base	RLOO $m = 2$	DAPO $m = 8$	POISE $m = 2$	Base	RLOO $m = 2$	DAPO $m = 8$	POISE $m = 2$
Code	63.33	69.37	68.49	**71.75**	25.02	62.54	61.41	**63.96**
Logic	7.14	22.95	**23.64**	22.28	9.53	13.31	15.14	**17.08**
Math	50.35	64.62	**68.85**	67.72	48.27	61.17	62.63	**64.02**
Simulation	61.71	62.25	61.46	**64.62**	46.58	58.17	55.33	**58.63**
Science	39.72	47.57	48.98	**50.18**	33.35	37.62	36.36	**39.69**
Tabular Reasoning	53.48	55.83	55.83	**57.09**	28.51	41.30	40.42	**42.84**
LiveBench	39.59	**45.78**	45.20	45.35	30.18	34.95	**36.59**	33.54
Instruction Following	**83.18**	80.78	79.30	81.70	78.56	76.34	**78.93**	77.26
Domain average	49.81	56.14	56.47	**57.59**	37.50	48.18	48.35	**49.63**

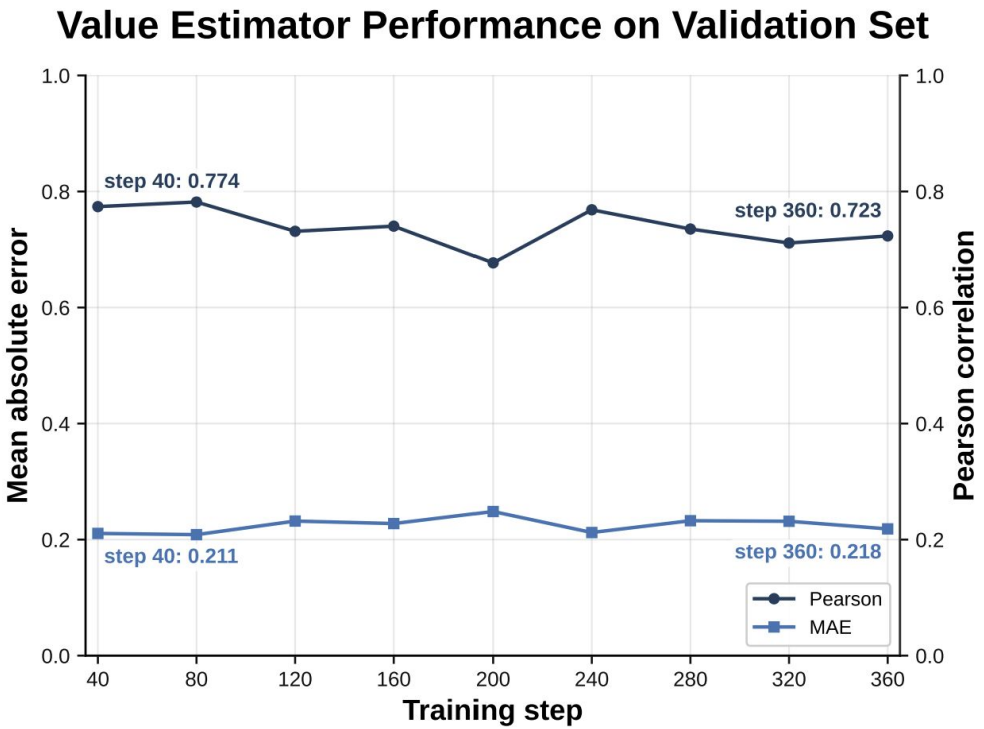
Analysis

Training Dynamics



Analysis

Training Dynamics



Analysis

Ablations: value estimation

Input feature	MAE ↓	r ↑
Prompt hidden states only	0.234	0.569
Reasoning hidden states only	0.132	0.821
Mean entropy only	0.152	0.780
Response length only	0.251	0.494
Full estimator	**0.126**	**0.838**

Analysis

Ablations: policy optimization

Method	Rollouts	Avg@32
base model	–	25.8
POISE (no trajectory-level feature)	1	47.5
POISE (no cross-rollout)	1	40.7
PPO (Token-level critic)	1	49.3
POISE (Ours)	2	**50.1**

Link to our paper



Contact us



Yunho Choi



Jongwon Lim