On Predictability of Reinforcement Learning Dynamics in Large Language Models

Yuchen Cai[1], Ding Cao[1], Xin Xu[3], Zijun Yao[2], Yuqing Huang[1], Zhenyu Tan[1], Benyi Zhang[1]
Guangzhong Sun[1], Guiquan Liu[1†],Junfeng Fang[2†]

[1]University of Science and Technology of China
[2]National University of Singapore
[3]Hong Kong University of Science and Technology

Email: caiyuchen@mail.ustc.edu.cn Code: https://github.com/caiyuchen-ustc/Alpha-RL

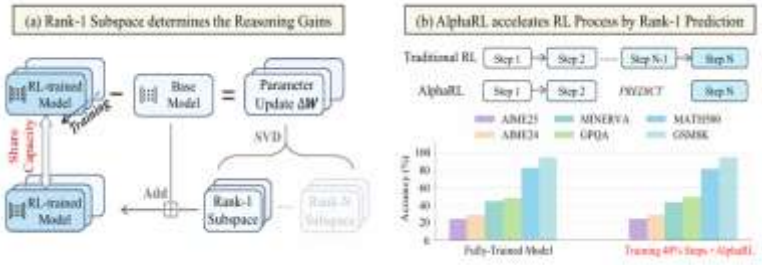
Introduction

Reinforcement learning (RL) has become a key driver behind the recent reasoning advances in large language models. However, the internal parameter dynamics during RL training remain largely unexplored, leaving the process as a black box.

In this work, we uncover two fundamental properties of RL-induced parameter updates in LLMs:

1.Rank-1 Dominance: The top singular subspace of the parameter update matrix ΔW captures over 99% of reasoning gains.

2.Rank-1 Linear Dynamics: This dominant subspace evolves in an approximately linear manner throughout training, enabling accurate prediction from early checkpoints.



Methods

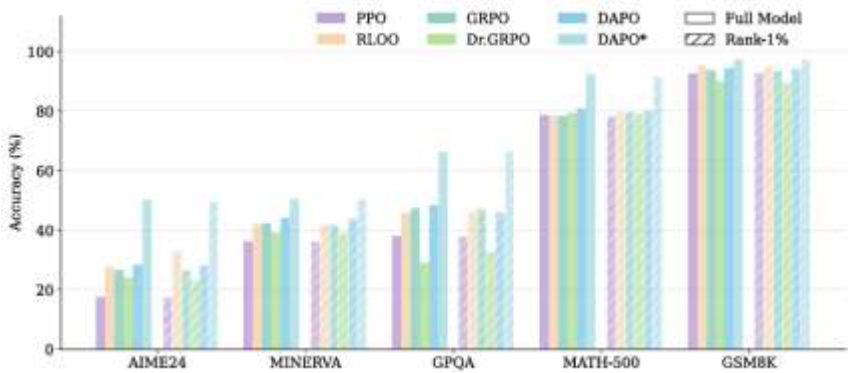
We describe the method for quantifying the contribution of the Rank-1 Subspace to reasoning gains in RL. Sepecifically, performing SVD on ΔW:

$$\Delta W = \sum_{i=1}^{r} \sigma_i u_i v_i^\top, \quad r = \mathrm{rank}(\Delta W).$$

$$\Delta W^{(1)} = \sigma_1 u_1 v_1^\top.$$

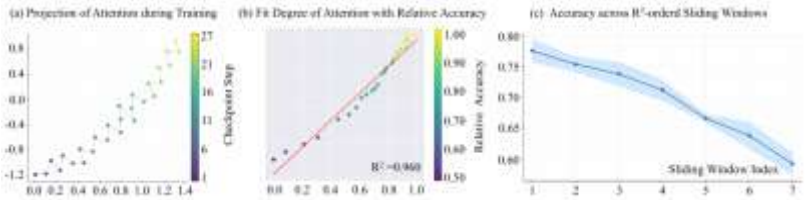
$$\Delta \hat{W}^{(1)} = \alpha \Delta W^{(1)}, \quad \alpha = \frac{\|\Delta W\|_2}{\|\Delta W^{(1)}\|_2}.$$

Key finding1: Rank-1 Dominance



Comparison between RL-trained models and their Rank-1(1%) parameter update counterparts across five reasoning benchmarks. The results demonstrate that retaining only the Top 1(1%) of the parameter update matrix is sufficient to recover the reasoning gains achieved by RL-trained models.

Key finding2: Rank-1 Linear Dynamics



Rank-1 Linear Dynamics reveals that the dominant singular subspace of RL-induced parameter updates evolves in an approximately linear manner throughout training, with an average R^2 exceeding 0.96. This property enables accurate prediction of final parameter updates from early training checkpoints, forming the foundation for acceleration methods like AlphaRL.

RL Accelerate framework: AlphaRL

Based on the two findings, we propose **AlphaRL**, which accelerates RL training by predicting the future updates of the Rank-1 subspace using only early-stage checkpoints:

$$y^{(t)} = \alpha z_1^{(t)} + \beta + \varepsilon^{(t)},$$

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{t=1}^{T} \left(y^{(t)} - (\alpha z_1^{(t)} + \beta)\right)^2,$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \left(y^{(t)} - \hat{y}^{(t)}\right)^2}{\sum_{t=1}^{T} \left(y^{(t)} - \bar{y}\right)^2}, \quad \hat{y}^{(t)} = \hat{\alpha} z_1^{(t)} + \hat{\beta}.$$

Stage	AIME24	AIME25	MATH	MINERVA	GPQA	GSM8K	Avg.
DAPO for the Qwen3-8B Base Model							
Fully Trained Model	**28.54**	**24.17**	**80.95**	**44.02**	48.23	94.35	**53.38**
Training 10%	12.50	7.50	70.25	32.07	36.66	84.30	40.55
Training 40%	15.80	11.67	77.60	37.07	41.67	93.20	46.30
Training 10%+AlphaRL	15.00	11.67	76.45	40.46	41.54	93.75	46.47
Training 40%+AlphaRL	28.33	23.75	80.50	43.27	**49.25**	**94.75**	53.31
GRPO from the Qwen3-8B Base Model							
Fully Trained Model	**26.40**	**21.67**	78.25	**42.19**	**47.10**	**93.50**	**51.52**
Training 10%	9.17	8.33	64.65	31.89	36.74	85.35	39.36
Training 40%	15.83	14.17	72.25	37.30	41.16	91.25	45.30
Training 10%+AlphaRL	12.50	13.25	67.60	36.83	36.74	91.35	43.43
Training 40%+AlphaRL	22.25	18.13	**78.45**	40.12	43.13	91.75	49.42
RLOO from the Qwen3-8B Base Model							
Fully Trained Model	**27.50**	18.33	**78.25**	**41.90**	**45.82**	**95.10**	**50.82**
Training 10%	11.67	8.33	57.25	35.02	38.65	83.50	39.89
Training 40%	16.67	14.17	72.75	39.24	42.05	93.75	46.44
Training 10%+AlphaRL	11.67	14.17	60.45	37.46	44.95	93.75	43.74
Training 40%+AlphaRL	17.92	**18.33**	76.00	40.60	44.40	93.80	48.52

Further Thoughts

Could the linear predictability of the Rank-1 subspace serve as a simple yet powerful signal for monitoring and stabilizing RL training — enabling early detection of anomalous updates or guiding adaptive interventions? **If you're interested in exploring this direction, feel free to reach out and discuss with me via email.**