NEURAL INFORMATION
PROCESSING SYSTEMS

Diversity-Aware Policy Optimization for Large Language Model Reasoning

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, Kay Chen Tan

Motivation

- While RL has been extensively applied to LLM reasoning, the role of diversity remains largely unexplored in this context, even though it plays a crucial role in RL research.

- Diversity plays a crucial role in RL research – can be generally divided into 3 categories
  - The first category uses diversity primarily to improve exploration efficiency, where diversity emerges as a byproduct of maximizing final task performance [1, 2, 3, 4, 5].
  - The second category treats diversity either as a constraint (optimizing quality subject to diversity constraints) or as an objective (optimizing diversity under quality constraints) [6, 7, 8, 9, 10].
  - The third category optimizes quality and diversity simultaneously, known as Quality-Diversity RL methods [11, 12, 13, 14].

- These findings naturally lead us to ask the following question: Is promoting diversity essential during RL training for LLM reasoning?
- In this work, our research tend to extend the first category on RL-based LLM training.

Motivation

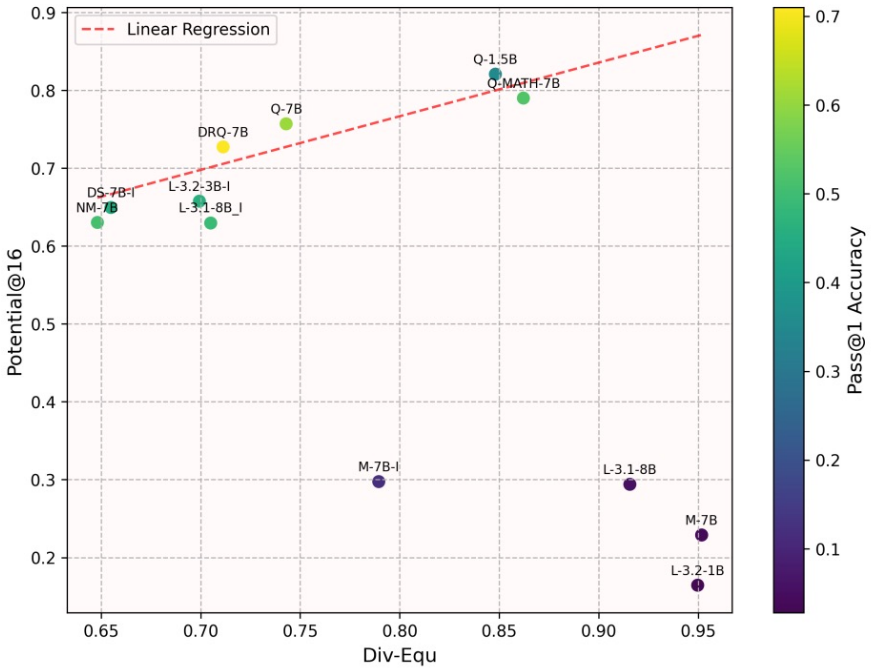
[1] Zhang-Wei Hong, et al. Diversity-driven exploration strategy for deep reinforcement learning. (NeurIPS 2018)
[2] Benjamin Eysenbach, et al. Diversity is all you need: Learning skills without a reward function. (arXiv)
[3] Jack Parker-Holder, et al. Effective diversity in population based reinforcement learning. (NeurIPS 2020)
[4] Edoardo Conti, et al. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. (NeurIPS 2018)
[5] Zhenghao Peng, et al. Non-local policy optimization via diversity regularized collaborative exploration. (arXiv, 2020)

[6] Muhammad A Masood and Finale Doshi-Velez. Diversity-inducing policy gradient: Using
396 maximum mean discrepancy to find a set of diverse policies. (arXiv, 2019)
[7] Yunbo Zhang, Wenhao Yu, and Greg Turk. Learning novel policies for tasks. (ICML, 2019)
[8] Tom Zahavy, Brendan O'Donoghue,et al. Discovering diverse nearly optimal policies with successor features. (arXiv 2021)
[9] Zihan Zhou, et al. Continuously discovering novel strategies via reward-switching policy optimization. (arXiv 2022)
[10] Mahsa Ghasemi, et al. Multiple plans are better than one: Diverse stochastic planning. (ICAPS 2021)

[11] Geoffrey Cideron, et al. Qd-rl: Efficient mixing of quality and diversity in reinforcement learning. (arXiv 2020)
[12] Sumeet Batra, et al. Proximal policy gradient arborescence for quality diversity reinforcement learning. (arXiv 2023)
[13] Thomas Pierrot, et al. Diversity policy gradient for sample efficient quality-diversity optimization. (GECC 2022)
[14] Bryon Tjanaka, et al. Approximating gradients for differentiable quality diversity in reinforcement learning. (GECC 2022)

Correlation between LLMs' reasoning potential and solution diversity

Potential-Diversity Experiment



$$\text{Potential@k} := \frac{\sum_{i=1}^{N} \text{Pass@k}(q_i) \cdot (1 - \text{Pass@1}(q_i))}{\sum_{i=1}^{N} (1 - \text{Pass@1}(q_i))},$$

$$\text{Div-Equ} := \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{U}_i|}{|\mathcal{A}_i|},$$

U_i, A_i: unique equations and all equations extracted from responses.

- Observation 1: For LLMs with low quality (accuracy), there is no obvious relationship between diversity and potential.
- Observation 2: For LLMs with high quality, there is generally a positive correlation between potential and diversity.

Since the optimization direction is guided by correct answers in multiple sampled responses, the result directly links our Potential@k metric to RL training improvements.

Entropy-based diversity

- A straightforward approach is to define diversity as the average entropy of the LLM's outputs per question.
- However, this formulation introduces length bias: longer responses inherently exhibit higher entropy.
- To address this issue, we introduce token-level entropy

$$\widehat{J}_{Div}(\pi_\theta) := \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{old}(\cdot|q)}\left[\frac{1}{T}\sum_{t=1}^{T}\mathcal{H}(\pi_\theta(\cdot|q,o^{<t}))\right],$$

- We further reformulate the diversity objective to enable effective backpropagation

$$\begin{aligned}\widehat{J}_{Div}(\pi_\theta) &= \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{old}(\cdot|q)}\left[-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\widetilde{o}^t\sim\pi_\theta(\cdot|q,o^{<t})}[\log\pi_\theta(\widetilde{o}^t|q,o^{<t})]\right]\\ &= \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{old}(\cdot|q)}\left[-\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_\theta(o^t|q,o^{<t})}{\pi_{old}(o^t|q,o^{<t})}\log\pi_\theta(o^t|q,o^{<t})\right].\end{aligned}$$

Promoting diversity on positive samples

- Directly applying the diversity objective in training will increase diversity in incorrect solutions. Intuitively, negative samples offer more room for diversity enhancement, which can skew the model's optimization process.
- To address this issue, concentrate on promoting diversity on positive samples:

$$J_{Div}(\pi_\theta) = \mathbb{E}_{q\sim\mathcal{Q}, o\sim\pi_{old}(\cdot|q)}\left[-\mathbb{I}(r=1)\cdot\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_\theta(o^t|q,o^{<t})}{\pi_{old}(o^t|q,o^{<t})}\log\pi_\theta(o^t|q,o^{<t})\right],$$

This is akin to fostering diversity in high-quality policies in population-based RL training [1]. Beyond this, we further justify this design by analyzing the gradient dynamics.

[1] Wu S, Yao J, Fu H, et al. Quality-similar diversity via population based reinforcement learning. (ICLR2023)

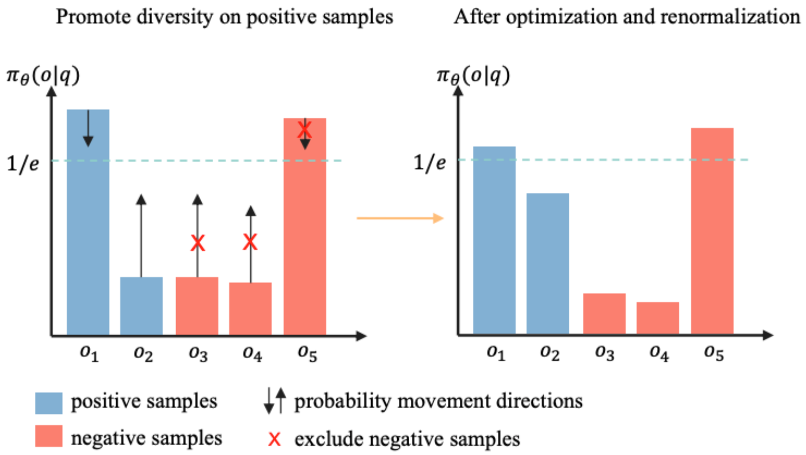
Promoting diversity on positive samples

- The gradient from the diversity objective:

$$\nabla_{\pi_\theta} \widehat{J}_{Div}(\pi_\theta) = \mathbb{E}_{q\sim\mathcal{Q},o\sim\pi_{old}(\cdot|q)}\left[-\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_\theta\left[\pi_\theta(o^t|q,o^{<t})\log\pi_\theta(o^t|q,o^{<t})\right]}{\pi_{old}(o^t|q,o^{<t})}\right].$$

$$-\nabla_\theta\pi_\theta(o^t|q,o^{<t})\log\pi_\theta(o^t|q,o^{<t}) = -[1+\log\pi_\theta(o^t|q,o^{<t})]\cdot\nabla_\theta\pi_\theta(o^t|q,o^{<t}).$$

- For tokens with small probs, the gradient tend to increase the probability.
- However, this tendency is undesirable for negative samples. Thus, excluding diversity enhancement for negative samples mitigates conflicts between solution quality and diversity.



Experiment Results

- Base models: Qwen2.5-Math-7B
- Benchmarks: GSM8K, MATH500, Olympiad Bench, and College Math

Promoting diversity can enhance the ability of LLM Reasoning.

Table 2: Avg@8 accuracy on mathematical benchmarks.

Method	GSM8K	MATH500	Olympiad Bench	College Math	Avg
Qwen2.5-Math-7B	53.37 (0.56)	48.10 (0.82)	15.80 (0.22)	19.36 (0.14)	34.16
R1-zero	87.77 (0.86)	72.97 (1.20)	37.26 (0.52)	42.22 (0.31)	60.06
R1-zero-Div (Ours)	**90.64 (0.89)**	**76.92 (1.24)**	**39.19 (0.55)**	**47.49 (0.32)**	**63.56**
SimpleRL-Zoo	89.46 (0.87)	77.15 (1.23)	39.43 (0.57)	47.19 (0.34)	63.31
Eurus-2-7B-PRIME	88.31 (0.86)	73.92 (1.18)	36.56 (0.50)	45.27 (0.30)	61.02

Our method can generate more diverse solutions.

Table 3: Diversity of different methods on GSM8K test set.

Method	Div-Equ	Div-N-gram	Div-Self-BLEU
Qwen2.5-Math-7B	92.26	29.29	85.98
Eurus-2-7B-PRIME	60.86	24.08	48.20
SimpleRL-Zoo	74.89	25.41	49.32
R1-zero	75.02	27.75	56.00
zero-Div (Ours)	**79.29**	**29.60**	**58.89**

Ablation Study

Analysis on the choice of diversity weights λ

Table 4: Ablation Study on different diversity weights on mathematical benchmarks

Method	GSM8K	MATH500	Olympiad Bench	College Math	Avg
$\lambda = 0$	88.7	74.6	37.3	43.3	61.0
$\lambda = 0.05$, pos	88.1	74.8	38.2	45.8	61.7
$\lambda = 0.02$, pos	<u>90.7</u>	<u>76.0</u>	<u>38.4</u>	<u>45.9</u>	<u>62.8</u>
λ= 0.01, pos	**91.7**	**78.2**	**40.1**	**47.6**	**64.4**
$\lambda = 0.01$, pos+neg	89.8	76.6	39.6	46.9	63.2

Experiment on 1.5B base model

Table 5: Ablation Study on Qwen2.5-Math-1.5B base model

Method	GSM8K	MATH500	Olympiad Bench	College Math	Avg
Qwen2.5-Math-1.5B	39.4	36.4	23.0	6.6	26.3
R1-zero	82.9	66.4	**32.1**	43.1	56.1
R1-zero-Div (Ours)	**83.2**	**70.4**	32.0	**43.9**	**57.4**