2026-02-15

2026 2 15 HuggingFace Papers + ArXiv

---

5 RLVR/GRPO

---

1. Unveiling Implicit Advantage Symmetry: Why GRPO Struggles with Exploration and Difficulty Adaptation

: [RLVR], [GRPO], [ ], [ ]

- **ArXiv ID**: 2602.05548
- : Zhiqi Yu, Zhangquan Chen, Mengting Liu, Heye Zhang, Liangqiong Qu
- : The University of Hong Kong (HKU)
- : 2026-02-05 (v2: 2026-02-12)

RLVR (Reinforcement Learning with Verifiable Rewards) LLM GRPO (Group Relative Policy Optimization) GRPO (exploration efficiency) (difficulty adaptation)

A_i i (advantage) GRPO Group Relative Advantage Estimation (GRAE)

$$A_i^{GRAE} = \frac{r_i - \mu_G}{\sigma_G}$$

r_i (=1 =0) μ_G σ_G

(Implicit Advantage Symmetry) 1. logits 2.

- GRPO
-

1. GRAE
2.
3. **Asymmetric GRAE (A-GRAE)**

- GRPO, PPO, DPO
- 7 benchmarks (LLM MLLM)
- LLM

- →
- →

1. ” ”
2. group size
3. p-values
4. **prior work** GRPO PPO/TRPO

RLVR —— + (symmetry breaking)
GRPO

---

2. Detecting RLVR Training Data via Structural Convergence of Reasoning

: [RLVR], [ ], [Benchmark], [ ]

- **ArXiv ID**: 2602.11792
- : Hongbo Zhang, Yue Yang, Jianhao Yan, Guangsheng Bao, Yue Zhang, etc.
- : Westlake University ( )
- : 2026-02-12

RLVR benchmark token RLVR

RLVR - RLVR prompt → rigid and similar - prompt →

- RLVR " " (structural convergence)
-

1. **Min-kNN Distance**
   - prompt completions
   - k
   - token (black-box)
2. RLVR

1. Min-kNN " = "
2. prompt
3. benchmarks
4. **prior work** membership inference

RLVR benchmark —— " " " "

---

3. Think Longer to Explore Deeper: Learn to Explore In-Context via Length-Incentivized Reinforcement Learning

: [RL], [In-Context Exploration], [Test-Time Scaling], [ ]

- **ArXiv ID**: 2602.11748
- : Futing Wang, Jianhao Yan, Yun Luo, Ganqu Cui, Zhi Wang, etc.
- : Westlake University
- : 2026-02-12

test-time scaling **In-Context Exploration**——

State Coverage theory —— **"Shallow Exploration Trap"** ( )

$p(L)$

$$p(L) \propto \exp(-\lambda L)$$

λ

1. State Coverage
2. **Length-Incentivized Exploration (LIE)**
   -
   - (redundancy penalty)
   -

- Qwen3, Llama
- in-domain + out-of-domain tasks
- In-domain +4.4% Out-of-domain +2.7%

1. LIE length reward exploration
2. **Redundancy Penalty** " "
3. **GRPO** A-GRAE
4. **Scaling**

test-time scaling " "—— (annealing) ( ) → ( )

---

4. dVoting: Fast Voting for dLLMs

: [ ], [ ], [Test-Time Scaling], [ ]

- **ArXiv ID**: 2602.12153
- : Sicheng Feng, Zigeng Chen, Xinyin Ma, Gongfan Fang, Xinchao Wang
- : National University of Singapore (NUS)
- : 2026-02-12

Diffusion Large Language Models (dLLMs) token test-time scaling

prompt sample - token - **token**

dVoting 1. samples 2. token 3. token 4.

Benchmark	
GSM8K	+6.22% ~ +7.66%
MATH500	+4.40% ~ +7.20%
ARC-C	+3.16% ~ +14.84%
MMLU	+4.83% ~ +5.74%

1. ” ”
2. “ ”
3. **Self-Consistency** Chain-of-Thought Self-Consistency
4.

dLLMs dLLMs test-time scaling ” ”——
” ”

---

5. ThinkRouter: Efficient Reasoning via Routing Thinking between Latent and Discrete Spaces

: [ ], [LLM], [ ]

- **ArXiv ID**: 2602.11683
- : Haoliang Wang, Xiang Chen, Tong Yu, etc.
- : UC San Diego + AI2
- : 2026-02-12

(latent space) (discrete space)

-
-

1. 6
2. **prior work** StreamLLM

1-2

	2024-2025	2026	
RLVR/GRPO	DeepSeek-R1, OpenReasoner	A-GRAE, LIE	→
	LLM Detection ( )	Min-kNN ( )	token
			→
dLLMs	MAR (Multi-Token)	dVoting	→
Test-Time Scaling	Self-Consistency, BoN	dVoting, LIE	→

RLVR

1. **A-GRAE** ( ) GRPO
2. **Min-kNN** ( ) RLVR
3. **LIE** ( ) In-Context Exploration
4. **dVoting** ( )

LLM ” - ”trade-off