ICLR 2026 Conference Presentation

Diversity-Enhanced Reasoning for Subjective Questions

MultiRole-R1

Yumeng Wang* · Zhiyuan Fan* · Jiayu Liu* · Jen-tse Huang · Yi R. (May) Fung

Hong Kong University of Science and Technology · Johns Hopkins University

Code & Data: github.com/yumeng-10/multirole-r1

+14.1%

In-domain accuracy gain

+7.64%

Out-of-domain gain

+5.78%

AIME 2024 gain

r = 0.74

Diversity-accuracy correlation

Motivation & Problem Statement

Large Reasoning Models (LRMs)

Excel at objective tasks:
Mathematical reasoning, code generation, commonsense QA

Trained via RLVR:
Reinforcement Learning with Verifiable Rewards

⚠️ **But RLVR degrades generation diversity!**

The Subjective Reasoning Gap

① No single ground truth — answers depend on role perspective

② RLVR's single-ground-truth optimization fails here

③ Existing diversity methods focus only on objective tasks

④ Multi-agent debate & prompting: no training-time solutions

→ **We propose MultiRole-R1: the first diversity-enhanced training framework for subjective reasoning**

MultiRole-R1: Two-Stage Framework

Goal: Diversify reasoning paths T for a subjective question Q using model M

STAGE 1

Perspective Diversity via SFT

① Generate n contrastive roles via few-shot prompting
② Sample k reasoning paths per role (temp τ = 1)
③ Self-consistency filtering via majority voting
④ Concatenate paths ⟶ Multi-Role CoT
⑤ Finetune model (SFT) on 2,700 synthesized entries

→

STAGE 2

Token-Level Diversity via GRPO

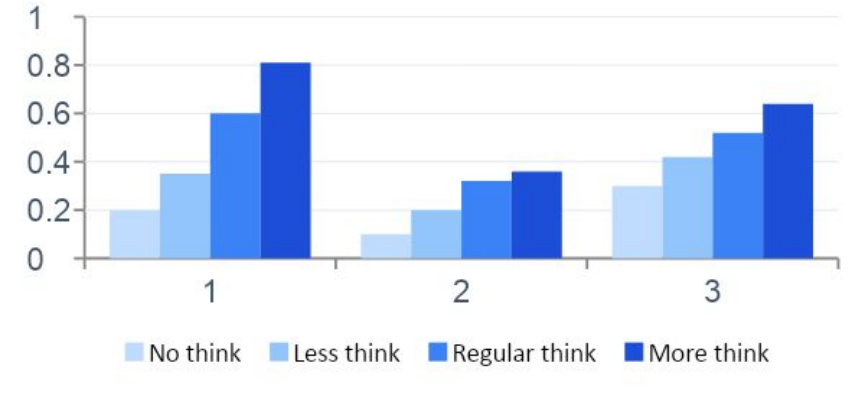
① GRPO: sample group of outputs, compare rewards
② Verifiable reward R_acc: role-based correctness
③ Diversity reward R_div: lexical + structural + discourse
④ Combined: $R = \delta \cdot R_{acc} + (1-\delta) \cdot R_{div}$
⟶ Maintains non-zero advantage, prevents diversity collapse

Pilot Analysis: Design Choices

Finding 1: Reasoning Length Scaling Law

More think significantly outperforms shorter settings on subjective tasks.

Performance peaks at **~3 'Wait' tokens** and diminishes beyond that point.

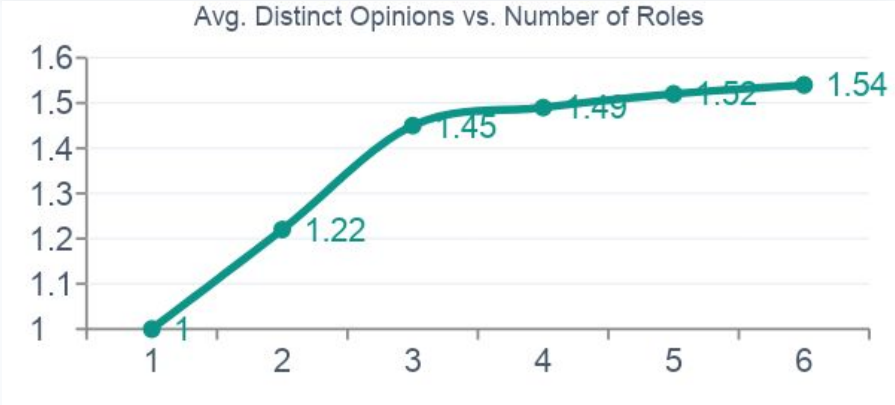


Finding 2: Role Perspective Scaling Law

Distinct opinions increase with more roles.

n = 3 roles is the turning point — beyond this, information gain diminishes.

→ **MultiRole-R1 uses 3 roles.**



Method: Merging Strategies & Diversity Reward

Merging Strategies (Stage 1)

Divergent Merging

For tasks with role-dependent answers (e.g., GLOQA, CALI). Each role's answer is compared to its own ground truth. Final prediction via weighted aggregation.

Convergent Merging

For tasks with a single correct answer (e.g., BBQ, ETHICS). Consensus reached via majority voting across role reasoning. Final answer = most-voted choice.

Diversity Reward R_div (Stage 2)

Combined score: $D_{final} = \Sigma \omega_i \cdot D_i$
(equal weights, 8 metrics)

Lexical (D_lex): Type-Token Ratio — vocabulary richness
Entropy (D_ent): Normalized token entropy
Length (D_len): Coefficient of variation of sentence lengths
Pattern (D_pat): Entropy over sentence types (declarative, etc.)
Adjacent (D_adj): Mean Jaccard distance between adjacent sentences
Yule's K: Vocabulary concentration (lower = more diverse)
Distinct N-gram: Proportion of unique n-grams
Function Word: Entropy over function-word distribution

Human Alignment: 3 PhD annotators rated 240 outputs (1–10 scale). Diversity metric achieves 0.88–0.95 alignment with human ratings.

Experimental Setup

Datasets

Training (In domain, Subjective)

BBQ, GlobalOpinionQA, ETHICS

Test Subjective

CALI (cultural NLI)

Test Objective

CSQA, GSM8K, AIME 2024

Total training

2,700 synthesized entries

Models Evaluated

7B

R1-Distill-Qwen-7B

8B

R1-Distill-Llama-8B

14B

R1-Distill-Qwen-14B

8B

Qwen3-8B (reasoning mode)

Baselines

ICL

Zero-Shot CoT, Self-Refine, Role-Play

Budget

More think (3× length)

SFT

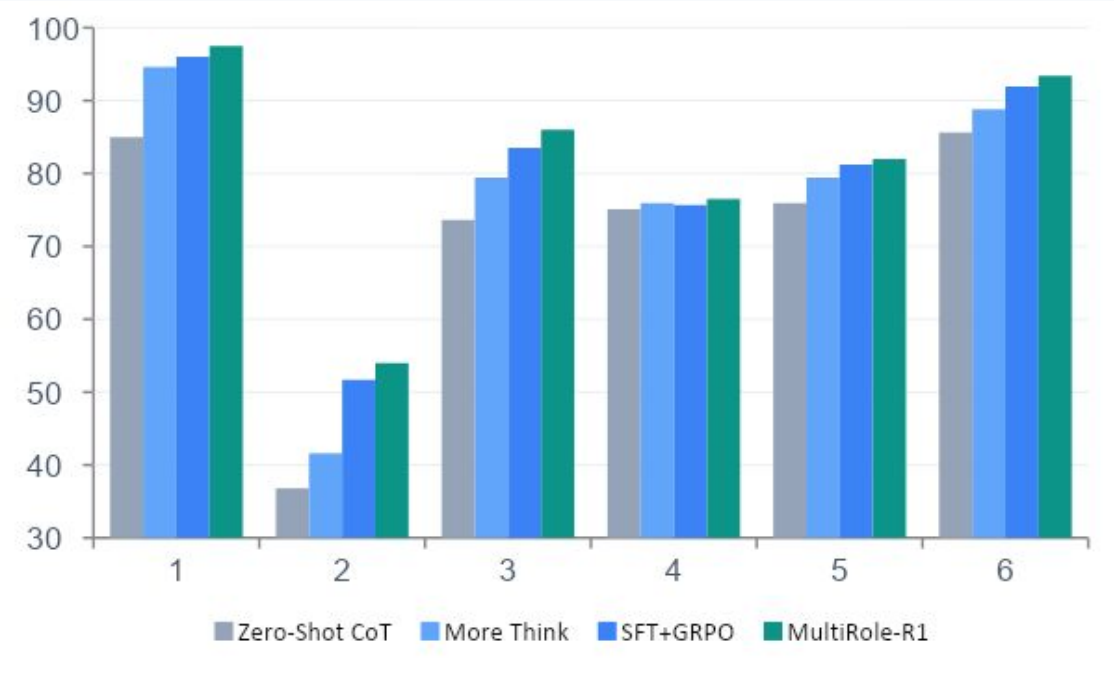
Self-consistency SFT only

RL

SFT + DPO, SFT + vanilla GRPO

Main Results

Accuracy Comparison — R1-Distill-Qwen-14B (representative model)



Key Numbers

+10.6% avg. accuracy over all tasks

+14.1% in-domain (BBQ, GLOQA, ETHICS)

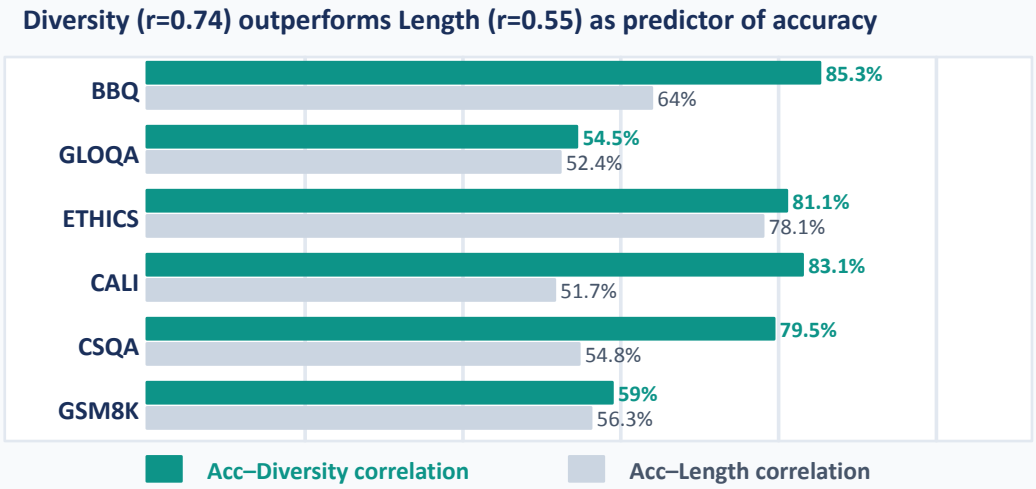
+7.64% out-of-domain tasks

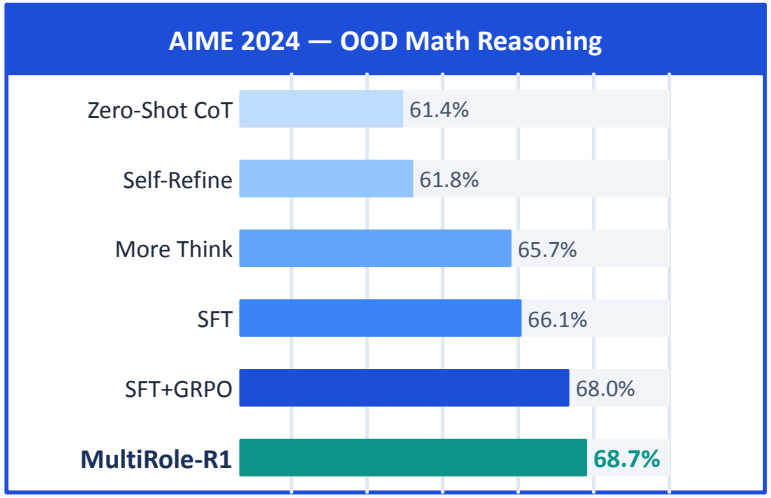
+5.78% AIME 2024 (unseen math)

+19.7% GRPO vs DPO accuracy gain

⚠️ On-policy GRPO (+19.7%) dramatically outperforms off-policy DPO (+2.4%): DPO's pairwise format cannot handle equally-valid subjective answers.

Analysis & Conclusion





Key Takeaways — MultiRole-R1

First diversity-enhanced training framework for subjective reasoning

SFT drives 8.3% of gain via multi-role perspective diversity

GRPO+RS drives 6.6% gain via token-level diversity reward shaping

Diversity is a stronger, more consistent predictor of accuracy than length

Generalizes to objective tasks including advanced math (AIME 2024)