

Diversity-Incentivized GRPO for Constrained Arithmetic Reasoning

Stanford CS224N Custom Project Mentored by Chenglei Si. No external collaborators, mentors or project sharing.

Gaurav Tyagi

Department of Computer Science
Stanford University
gtyagi@stanford.edu

Abstract

Reinforcement learning with verifiable rewards (RLVR) improves LLM reasoning, but sparse rewards and standard policy optimization may cause models to converge onto a single solution strategy even when many valid approaches exist. This project investigates whether adding a structural diversity incentive to Group Relative Policy Optimization (GRPO) increases the variety of correct reasoning paths without hurting accuracy [1]. We test this on the Game of 48, a constrained arithmetic reasoning task where five integers must be combined with $+$, $-$, $*$, $/$ and parentheses to equal 48. We fine-tune Qwen3-8B instruction-tuned and base models under three conditions—no fine-tuning, standard GRPO ($\lambda = 0$), and diversity-augmented GRPO ($\lambda = 1$)—and evaluate on 405 held-out problems using Pass@1, Pass@ k , Unique Correct@ k , and Avg Symbolic Diversity metrics. We find that standard GRPO significantly improves correctness but *reduces* symbolic diversity relative to the zero-shot baseline, confirming mode collapse. Diversity-augmented GRPO recovers and substantially exceeds zero-shot diversity while sustaining most of the correctness gains, achieving up to 53% higher Avg Symbolic Diversity and 15% more unique correct solutions than the GRPO baseline, with only a modest reduction in Pass@1.

1 Introduction

Large language models trained with reinforcement learning from verifiable rewards (RLVR) have shown impressive gains on mathematical reasoning benchmarks. Methods such as GRPO [?] update the policy by comparing a group of rollouts against each other, reinforcing high-reward solutions and discouraging low-reward ones. While effective at improving correctness, these methods suffer from a well known failure mode: once the model discovers a reliable solution strategy, it exploits it exclusively, collapsing onto a single reasoning pathway. This mode collapse is problematic because it limits the model’s ability to generalize to problem variations that require different solution structures.

This project asks a focused empirical question: *can adding a group-level structural diversity bonus to the GRPO reward prevent mode collapse and increase the variety of valid reasoning paths, without sacrificing single-sample accuracy?* We examine this question on the Game of 48, a combinatorial arithmetic task well-suited to studying diversity because most problem instances admit multiple structurally distinct correct solutions. In addition, solution validity is extremely easy to calculate. Our diversity metric operates at the level of canonical expression structure (abstract syntax tree form), ensuring that trivially equivalent expressions (e.g. $A + B$ and $B + A$) are not double-counted.

Our main contributions are:

1. A diversity-augmented GRPO reward that only adds the diversity bonus for correct responses, preventing reward hacking with incorrect but diverse outputs.

2. An AST-based canonicalization pipeline that defines structural equivalence precisely and is used both to filter the training dataset and to score diversity at training and evaluation time.
3. An ablation study across six conditions (two model variants $\times$ three fine-tuning regimes) demonstrating that standard GRPO causes mode collapse while the diversity reward substantially reverses it.

2 Related Work

RLVR and GRPO. Reinforcement learning from verifiable rewards has emerged as a powerful paradigm for improving LLM reasoning without requiring human preference annotations [?]. GRPO computes advantages within a group of rollouts for the same prompt, avoiding the need for a separate value network. Recent work including DeepSeek-R1 and related models has demonstrated that GRPO can yield large gains on mathematical benchmarks [?].

Mode collapse and diversity in RLVR. A known failure mode of policy gradient methods applied to LLMs is exploitation of a narrow solution strategy [2]. Several works address this through entropy regularization or stochastic token selection [3, 4], which improve exploration locally but do not directly incentivize global diversity of reasoning paths.

DIVER. The most directly related work is DIVER (Diversity-Incentivized Exploration for Versatile Reasoning) by Hu et al. [1], which introduces two group-level diversity metrics—Textual Diversity (inverse BLEU between responses) and Equational Diversity (fraction of unique formulas not shared with other group members)—and integrates them as intrinsic rewards into GRPO. DIVER demonstrates that diversity incentives improve $\text{Pass}@k$ and out-of-domain generalization on several math benchmarks. Our work differs in three ways: (1) we use a stricter AST canonicalization-based diversity metric that eliminates commutative and associative redundancy; (2) we test on a combinatorial arithmetic task (Game of 48) where ground-truth solution diversity can be precisely and quickly calculated; and (3) we include a base model ablation to test whether instruction tuning interacts with the diversity reward.

Synthetic verifiable datasets. Our dataset generation approach is inspired by SynLogic [5], which provides infrastructure for generating and verifying Game of 24 instances. We adapt this to the five-number, target-48 variant.

3 Approach

3.1 Task and Dataset Generation

The **Game of 48** is defined as follows: given five integers drawn from $[1, 13]$, produce an arithmetic expression using each number exactly once with $+$, $-$, $*$, $/$ and parentheses that evaluates to exactly 48. We generate a synthetic dataset using a recursive symbolic solver that enumerates all structurally distinct valid expressions for a given hand. We keep only hands with at least **8 structurally distinct valid solutions**, ensuring every training example has sufficient diversity for the reward signal to be informative. The final dataset contains 5,000 training problems and 405 test problems with no overlap.

3.2 Verification and Canonicalization

Correctness is verified with a safe, AST-based symbolic checker that (1) parses the expression, (2) confirms the exact multiset of input numbers is used, and (3) evaluates the expression while catching division by zero and other illegal operations. This avoids the security risks of `eval()`-based checkers.

For diversity, we canonicalize expressions by flattening associative chains and sorting the operands of commutative operators ($+$ and $*$), so that $A + B$ and $B + A$ map to the same canonical form. Two expressions are considered *structurally equivalent* if and only if their canonical AST representations are identical.

3.3 Diversity-Augmented GRPO

We fine-tune models with GRPO and augment the reward with a structural diversity bonus:

$$R(o_i) = c_i + \lambda \cdot c_i \cdot d_i \quad (1)$$

where $c_i \in \{0, 1\}$ is the binary correctness indicator and

$$d_i = \frac{1}{G-1} \sum_{j \neq i} \mathbf{1}[\text{canonical}(o_i) \neq \text{canonical}(o_j)] \quad (2)$$

is the fraction of other rollouts in the group of size G whose canonical form differs from o_i . The c_i gate is crucial: it ensures the diversity bonus is only awarded for *correct* responses, preventing the model from being rewarded for generating incorrect-but-diverse outputs.

Training loop. Fine-tuning is implemented on top of the Tinker remote training framework [6], which hosts the model weights and GPU computation behind a client API. Tinker always applies LoRA adapters to the base model and exposes three operations used per training step: (1) `save_weights_and_get_sampling_client`, which snapshots the current adapter weights and returns a sampling client; (2) `forward_backward`, which runs the importance-sampling GRPO loss on a batch of Datum objects (each containing token sequences, reference log probabilities, and per-token advantages); and (3) `optim_step`, which applies an Adam update to the LoRA parameters.

Our training code orchestrates the outer GRPO loop. For each batch of P prompts, we dispatch G parallel sampling requests per prompt to Tinker and collect the resulting token sequences and log probabilities. We then compute rewards and advantages locally—this is where the diversity metric is applied—and package each rollout into a Datum for the forward-backward call. Advantages are computed within each group by subtracting the group mean reward; groups where all advantages are zero (all responses correct or all wrong with identical scores) are skipped to avoid degenerate gradient steps.

3.4 Baselines

We evaluate three fine-tuning conditions for both an instruction-tuned model (Qwen3-8B) and a base model (Qwen3-8B-Base):

- **No fine-tuning**: zero-shot performance of the pre-trained model.
- **Standard GRPO** ($\lambda = 0$): correctness reward only.
- **DIVER** ($\lambda = 1$): correctness plus diversity bonus.

4 Experiments

4.1 Data

We use the Game of 48 dataset described above: 5,000 training problems and 405 test problems, each requiring at least 8 structurally distinct correct solutions. This makes it theoretically possible for the model to generate 8 unique solutions in a size 8 rollout. The dataset generation code is adapted from the SynLogic repository [5], which provides analogous infrastructure for Game of 24. We note that generating 500 test problems proved infeasible under the 8-solution filter after 200,000 sampling iterations; relaxing this constraint (e.g., requiring only 4 distinct solutions) would allow a larger test set at the cost of reducing the max potential diversity of responses.

4.2 Evaluation Metrics

We sample $k = 8$ responses per test problem at temperature 1.0 and compute:

- **Pass@1**: fraction of problems where the first sampled response is correct.
- **Pass@ k** : fraction of problems where at least one of k responses is correct.
- **Unique Correct@ k** : average number of structurally distinct correct canonical forms found across k samples per problem.

- **Avg Symbolic Diversity:** for each problem, the fraction of correct-solution pairs (i, j) with $i \neq j$ whose canonical forms differ, averaged across all problems.

Pass@1 and Pass@ k measure correctness; Unique Correct@ k and Avg Symbolic Diversity jointly measure the breadth of the model’s solution repertoire. Importantly, Avg Symbolic Diversity is zero for any problem where fewer than two correct solutions are found, so it is sensitive to both the correctness rate and the structural variety of correct responses.

4.3 Experimental Details

Setting	Value
Models	Qwen3-8B, Qwen3-8B-Base
Fine-tuning framework	Tinker (LoRA, rank 32) [6]
Training problems	5,000
Training epochs	8
Batch size (P)	64
Group size (G)	8
Learning rate	10^{-6}
λ (DIVER run)	1.0
Max tokens	1,024
Training steps	~ 625 (8 epochs $\times$ 78 steps/epoch)

Table 1: Training configuration for all fine-tuned runs.

All fine-tuned models use identical hyperparameters; the only difference between conditions is the value of λ . Tinker applies LoRA adapters targeting all attention projection matrices (query, key, value, output) and all MLP projection matrices (gate, up, down) at rank 32, with $\alpha = 64$. Tinker’s importance-sampling loss re-weights each token’s policy gradient by the ratio of current to reference log probabilities, implementing standard off-policy GRPO. The Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\varepsilon = 10^{-8}$) is applied after each batch of $P = 64$ prompts. Training runs 8 epochs over the 5,000-problem training set, giving ~ 625 total gradient steps and substantial exposure to the diversity reward signal.

4.4 Results

Model	Condition	Pass@1	Pass@8	Unique@8	Avg Div
Qwen3-8B	No fine-tuning	0.295	0.593	1.027	0.310
Qwen3-8B	GRPO ($\lambda = 0$)	0.440	0.848	1.220	0.270
Qwen3-8B	DIVER ($\lambda = 1$)	0.417	0.780	1.405	0.412
Qwen3-8B-Base	No fine-tuning	0.250	0.576	0.930	0.280
Qwen3-8B-Base	GRPO ($\lambda = 0$)	0.420	0.810	1.205	0.272
Qwen3-8B-Base	DIVER ($\lambda = 1$)	0.402	0.772	1.431	0.437

Table 2: Evaluation results on 405 test problems, $k = 8$ samples per problem. Unique@8 = Unique Correct@ k ; Avg Div = Avg Symbolic Diversity.

5 Analysis

Standard GRPO causes mode collapse. The most striking result is that standard GRPO ($\lambda = 0$) *reduces* Avg Symbolic Diversity relative to the zero-shot baseline on both model variants (0.270 vs. 0.310 for the instruct model; 0.272 vs. 0.280 for the base model). This confirms the mode collapse hypothesis: as the model learns a reliable solution strategy, GRPO’s advantage normalization reinforces that strategy to the exclusion of alternatives. The correctness gains are substantial—Pass@1 improves by $\sim 49\%$ for the instruct model—but they come at the cost of solution variety.

DIVER substantially reverses mode collapse. Diversity-augmented GRPO recovers and exceeds the zero-shot diversity baseline: Avg Symbolic Diversity reaches 0.412 (instruct) and 0.437 (base), compared to 0.310 and 0.280 at zero-shot. Relative to the GRPO baseline, DIVER improves Avg Symbolic Diversity by 53% (instruct) and 61% (base). Unique Correct@8 similarly improves by 15% (instruct) and 19% (base) over GRPO. These gains confirm that the diversity bonus successfully incentivizes the model to explore structurally distinct solution pathways within each rollout group.

Diversity gains trade off against pass rate. DIVER’s diversity gains come with a modest correctness cost: Pass@1 drops by 0.023 (instruct) and 0.018 (base) relative to GRPO, and Pass@8 drops by 0.068 and 0.038 respectively. This suggests a tension between exploitation (repeatedly applying a known-correct strategy) and exploration (discovering structurally distinct approaches). The trade-off is moderate: DIVER still substantially outperforms no fine-tuning on both correctness metrics, and the absolute Pass@1 values (0.417, 0.402) remain competitive.

Base vs. instruct models. The base model shows a slightly stronger diversity effect than the instruct model under DIVER: higher Unique Correct@8 (1.431 vs. 1.405) and higher Avg Symbolic Diversity (0.437 vs. 0.412). One possible explanation is that the instruction-tuned model has already been shaped by RLHF to favor a small set of high-confidence response patterns, making it harder to diversify further. The base model, lacking this prior, responds more flexibly to the diversity reward. Notably, both models show similar correctness after fine-tuning, suggesting GRPO is effective at instilling the task objective regardless of starting point.

Interpretation of Unique Correct@8. Across all fine-tuned conditions, Unique Correct@8 is between 1.2 and 1.4, meaning fine-tuned models find on average fewer than 1.5 structurally distinct correct solutions per problem in 8 tries. Given that the dataset filters for problems with at least 8 distinct solutions, there is substantial headroom for improvement. The diversity reward nudges the model toward broader exploration, but the effect is incremental rather than transformative under the current $\lambda = 1$ setting. Larger λ , longer training, or curriculum strategies may be needed to fully exploit the available solution diversity.

6 Conclusion

We investigated whether a group-level structural diversity bonus can prevent mode collapse in GRPO fine-tuning without sacrificing correctness, using the Game of 48 as a testbed. Our main findings are: (1) standard GRPO causes measurable mode collapse, reducing symbolic diversity below the zero-shot baseline; (2) diversity-augmented GRPO (DIVER) substantially reverses this, achieving 53–61% higher Avg Symbolic Diversity and 15–19% more unique correct solutions than GRPO, at the cost of a modest ($\sim 5\%$) reduction in Pass@1; and (3) this pattern holds for both instruction-tuned and base model variants, with the base model showing a slightly larger diversity response.

These results support the hypothesis that explicit diversity incentives in RLVR are necessary to prevent policy collapse onto a single solution mode, and that the correctness gate in the reward function ($c_i \cdot d_i$) is sufficient to prevent reward hacking. Future work should explore higher values of λ , larger group sizes (which would provide a richer diversity signal per gradient step), and extension to more complex reasoning tasks where structural diversity has broader implications for generalization.

References

- [1] Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. Diversity-incentivized exploration for versatile reasoning, 2025.
- [2] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *CoRR*, abs/1702.08165, 2017.
- [3] Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning, 2025.
- [4] Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism, 2025.

- [5] Junteng Liu, Yuanxiang Fan, Zhuo Jiang, Han Ding, Yongyi Hu, Chi Zhang, Yiqi Shi, Shitong Weng, Aili Chen, Shiqi Chen, Yunan Huang, Mozhi Zhang, Pengyu Zhao, Junjie Yan, and Junxian He. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond, 2025.
- [6] Thinking Machines Lab. Tinker, 2025.