Extended Abstract

Motivation This paper seeks to increase both zero-shot and finetuned performance of LLMs on math-based tasks by increasing diversity. We test two hypotheses: for zero-shot, "Does requesting creative responses increase performance?", and for finetuned, "Does rewarding diversity directly increase performance?"

Implementation To implement the reinforcement-learning-based finetuning for this project, I relied on HuggingFace's *trl* package and implementation of group relative policy optimization/dynamic sampling policy optimization (GRPO/DAPO).

Results Higher temperatures degraded baseline Qwen2.5-Math-* performance on the GSM8K testing set, though the effect is less prominent with the largest model. Prompting for creativity did noticably increase zero-shot performance for all Qwen2.5-Math models. However, both prompting for creativity and rewarding diversity directly damaged finetuned performance.

Discussion By recreating baselines, we see both how finetuning naturally increases task-specific performance, and also how high temperatures may not be useful for math reasoning tasks. Additionally, we see that, across the board, prompting for creative responses increases zero-shot performance when writing solutions to math problems.

Conclusion This paper proposes two methods to improve LLM chain-of-thought-based performance on math tasks using the assumption that diverse solution spaces allow for better mathematical reasoning generalization. The first method is to prompt for creativity. This showed consistent zero-shot performance increases across all model sizes. Applications of CCoT include increasing inference performance with low-compute. The second method is to reward diversity directly. This paper shows performance degradation when rewarding diversity, though this may be due to a confluence of experimental design issues.

Begging for DEI: Increasing LLM Math Performance by Increasing Diversity

Tyler Ho
Department of Computer Science
Stanford University
tylerho@stanford.edu

Abstract

Large language models' (LLMs) zero-shot performance on math tasks depend heavily on the model size. This paper proposes two methods to increase both zero-shot performance and post-training performance on the GSM8K dataset: requesting creative responses and rewarding diversity directly. When requesting creative responses, all sizes of Qwen2.5-Math performed significantly[1] better. However, rewarding diversity directly through finetuning through group relative policy optimization-like (GRPO) methods did not increase performance and may have degraded task-specific performance.

1 Introduction

Large language models (LLMs) have clearly irreversibly changed our society and have shown remarkable zero-shot and few-shot learning capabilities. However, one sector in which LLMs are not able to learn well is mathematics. LLMs primarily generate tokens that may have some level of inherent meaning in their latent spaces. However, without concrete semantic meaning, LLMs struggle with creating correct solutions to math equations, where numbers' concrete meaning is the primary driver of correctness.

Rather than directly generating the next token to a fill-in-the-blank equation, researchers extended LLMs' ability to reason by having the models structure their mathematical reasoning as humans would. Chain-of-thought reasoning is an LLM generation technique in which an LLM writes human-like logic to approach a final answer (Wei et al., 2023). With LLMs' reasoning made transparent, researchers have found that LLMs' solution space for any given problem is small. A larger solution space may allow LLMs to solve new problems or produce novel solutions to existing problems. In addition, rewarding diverse but correct responses may improve performance on reasoning-heavy tasks.

The central hypotheses of this project is twofold: prompting for creativity can expose latent reasoning diversity in a base model and increase mathematical performance, and using reinforcement learning (RL) that rewards diversity can increase performance against correctness-only RL. Unlike the latter, which may collapse toward common solution templates, a diversity-aware reward may preserve multiple valid reasoning strategies while improving accuracy.

2 Related Work

DeepSeek-R1 relies on reinforcement learning and high-temperature sampling to encourage exploration during reasoning (Guo et al., 2025). Although this may allow models to create correct solutions

[1] Due to compute constraint, I was not able to do statistical significance tests.

with diverse vocabulary, the researchers did not reward the model for finding multiple distinct correct solutions. As a result, this may lead to models that converge to solution patterns that previously received high rewards.

Reinforcement learning from human feedback, or RLHF, may allow for unique responses while also verifying correctness, but the cost of human labeling makes it difficult to scale (Christiano et al., 2017). Direct preference optimization (Rafailov et al., 2024), or DPO, addresses the issue of human-made labels by allowing models to learn from preference-based signals. However, as a result, DPO can still produce models whose solutions resemble those that were rewarded in the past.

Self-consistency and best-of-n sampling also improve mathematical reasoning by generating multiple solutions and selecting an answer from them (Wang et al., 2023). However, these methods mainly use diversity at inference time rather than in training. This project differs by not only using prompt-induced diversity, but also rewarding diversity directly.

3 Method

3.1 Training Algorithm

Researchers at Deepseek propose a technique called group relative policy optimization (GRPO) in which multiple possible solutions are sampled with each prompt, and the model receives a grouped reward based on those responses (Shao et al., 2024). By giving rewarding the group, the model is able to learn which trajectories are good and which are not. In addition, the model is able to try multiple paths to find the correct solution.

To calculate GRPO loss, let $\mathbb{D}_{\mathrm{KL}}[\pi_\theta \mid\mid \pi_{\mathrm{ref}}]$ denote KL divergence (Shao et al., 2024):

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(r)}{\text{std}(r)}$$

$$r_{\pi_\theta} = \frac{\pi_{\text{ref}}(o_{i,t} \mid q, o_{i<t})}{[\pi_\theta(o_{i,t} \mid q, o_{i<t})]_{\text{no grad}}}$$

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left\{r_{\pi_\theta}\hat{A}_{i,t}, \text{clip}\left(r_{\pi_\theta}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right\} - \beta\mathbb{D}_{KL}[\pi_\theta \mid\mid \pi_{\text{ref}}].$$

Though GRPO was revolutionary in reinforcement learning with verifiable rewards (RLVR), it under-penalized longer responses. To address this limitation, a team at ByteDance propose the Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) algorithm (Yu et al., 2025). DAPO loss is calcluated as follows:

$$\mathcal{L}_{\text{DAPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{\pi_\theta}\hat{A}_{i,t}, \text{clip}\left(r_{\pi_\theta}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right) - \beta\mathbb{D}_{KL}[\pi_\theta \mid\mid \pi_{\text{ref}}].$$

This project proposes two methods for increasing LLM-generated solutions to math problems:

1. Request the model be creative directly (prompting)
2. Use DAPO to reward diversity directly (reward functions)

- Chain-of-Thought (CoT):

  Please reason step by step, and put your final answer within \boxed{}.

- Creativity Prompting Chain-of-Thought (CCoT):

  Solve the math problem below step by step. Show the key reasoning clearly, but keep the solution concise. Be creative in your approach: consider elegant shortcuts, alternative methods, visual or conceptual insights, or non-obvious simplifications when useful. Still prioritize correctness and clarity over novelty. At the end, provide the final answer on its own line in exactly this format:
  \[\boxed{<final answer>} \]
  Do not put any extra text after the boxed final answer.

Let o_i be responses. If o_i is correct, let c_i denote the Qwen3-Embedding-0.6B response embedding, and let C be the set of correct solution embeddings. Finally, let λ denote the diversity scalar.

- Diversity-Unaware:

$$r = \sum_{i=1}^{n} 1[o_i \text{ is correct}]$$

- Diversity-Aware:

$$b = \begin{cases} 0 & |C| \leq 1 \\ \lambda \min_{i \neq j} \text{cosine difference}(c_i, c_j) & \text{otherwise} \end{cases}$$

$$r = \sum_{i=1}^{n} 1[o_i \text{ is correct}] + b$$

4 Experimental Setup

First, I define the baselines as the performance when using the default CoT prompt provided by Qwen2.5-Math on all sizes (1.5B, 7B, and 72B). To test the hypotheses, I finetuned Qwen2.5-Math-1.5B on the GSM8K training set for two epochs using DAPO loss with the following hyperparameters.

The generation parameters were:

Max Tokens	Temperature	Top p	# of Generations
512 / 1024*	1.2 / 0.6*	0.9	4 / 8*

Table 1: Generation Control Variables. $*$: Hyperparameters used in the lower temperature baseline.

The training parameters were:

Batch Size	# of Epochs	LoRA Rank	LoRA α	LoRA Dropout	Similarity Scalar λ	KL β	Clip ε
128	2	16	32	0.05	1.0	0.04	0.2

Table 2: Training Control Variables

The decision variables are the system prompts and reward functions. This gives us 4 experimental setups: {(CoT, $\mathrm{num\_correct}$), (CCoT, $\mathrm{num\_correct}$), (CoT, $\mathrm{diversity}$), (CCoT, $\mathrm{diversity}$)}

4.1 Dataset

To train and test this method, I used the GSM8K dataset presented by OpenAI (Cobbe et al., 2021). It is a dataset of grade-school math questions, written in a "linguistically diverse" manner, with natural language solutions (Cobbe et al., 2021). Using Hugging-Face's datasets, the training set has 7.47k examples and the test set is 1.32k problems. I measured performance on math questions based on the testing split of OpenAI's grade school math dataset (gsm8k). In particular, I chose performance on the pass@1 and pass@4 metrics, in which the model is assessed on its ability to generate a correct answer within 1 or 4 generations, respectively.

The decision variables are the system prompts and reward functions.

5 Results

5.1 Quantitative Evaluation

Baseline performance is defined as zero-shot performance on the GSM8K test set against the pass@k metrics for $k \in \{1, 4\}$. As a reminder, CoT refers to chain-of-thought and CCoT refers to creative

chain-of-thought, where creative chain-of-thought is creativity requesting system prompt. I used a higher temperature with less generated tokens and a lower temperature with more generated tokens.

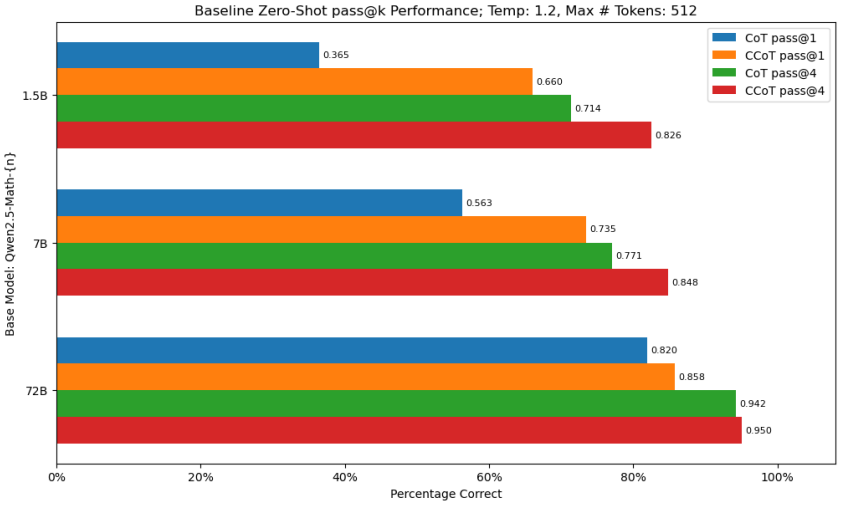

Figure 1: High Temperature Baseline Performance

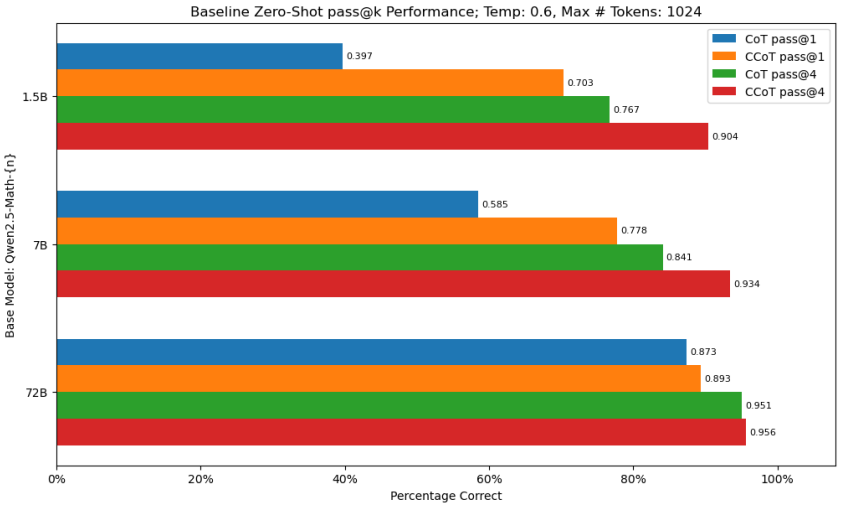

Figure 2: Low Temperature Baseline Performance

The higher temperature baseline overall performed worse than the lower temperature baseline, which is consistent with the general understanding of temperature on mathematical problems. Additionally, we see that prompting for creativity consistently increases zero-shot performance irrespective of model size. This is a new finding that contradicts the initial finding during the poster session where the creative prompt was shown to degrade performance on the 72B parameter Qwen model.

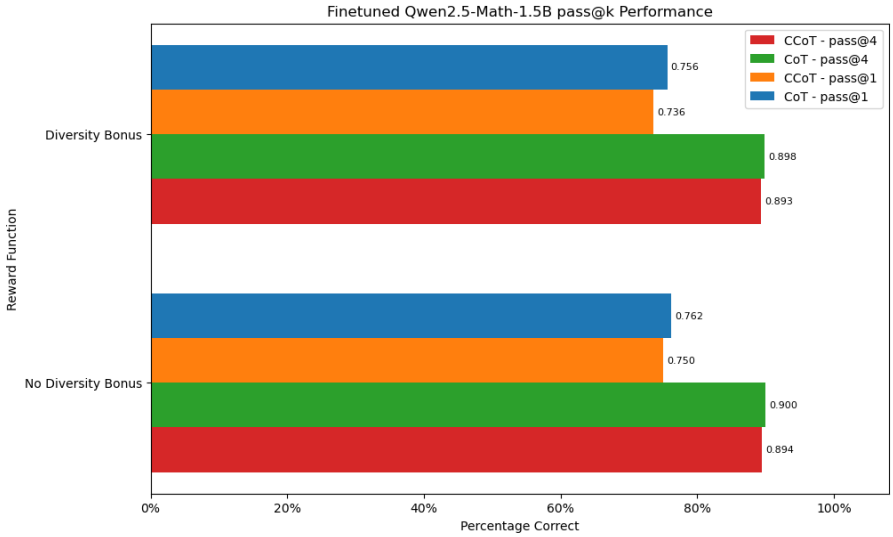

Figure 3: Experiment Results

Naturally, we see finetuning did increase the competitiveness of Qwen2.5-Math-1.5B's performance against larger models' zero-shot performance. However, rewarding diversity directly seems to mildly degrade performance. This may be due to the similarity scalar being too small. With the similarity scalar at 1, the diversity reward may have added noise rather than a clear signal. Oddly enough, we also see CCoT generally degrades performance on all metrics.

6 Discussion

By recreating baselines, we see both how finetuning naturally increases task-specific performance, and also how high temperatures may not be useful for math reasoning tasks. Additionally, we see that, across the board, prompting for creative responses increases zero-shot performance when writing solutions to math problems.

There are some major holes in the experimental rigor that are important to mention. Due to compute constraints, I was not able to finetune Qwen2.5-Math-1.5B using the baseline lower temperature settings for a fair apples to apples comparison. For the same reason, I was not able to measure reasoning path divergence, as defined in (Ju et al., 2025), of the models to compare each model's solution diversity. Similarly, I was not able to evaluate finetuned performance using a similarity scalar that is proportional to the number of generationed responses.

7 Conclusion

This paper proposes two methods to improve LLM chain-of-thought-based performance on math tasks using the assumption that diverse solution spaces allow for better mathematical reasoning generalization. The first method is to prompt for creativity. This showed consistent zero-shot performance increases across all model sizes. Applications of CCoT include increasing inference performance with low-compute. The second method is to reward diversity directly. This paper shows performance degradation when rewarding diversity, though this may be due to a confluence of experimental design issues.

8 Team Contributions

- **Tyler Ho: All**

Changes from Proposal Originally, this project was motivated to increase diversity in LLM-generated solutions to math problems. However, as it grew, I became more focused on increasing zero-shot performance rather than focusing on diversity specifically.

References

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] `https://arxiv.org/abs/2110.14168`

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

Feng Ju, Zeyu Qin, Rui Min, Zhitao He, Lingpeng Kong, and Yi R Fung. 2025. Reasoning Path Divergence: A New Metric and Curation Strategy to Unlock LLM Diverse Thinking. *arXiv preprint arXiv:2510.26122* (2025).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] `https://arxiv.org/abs/2305.18290`

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] `https://arxiv.org/abs/2402.03300`

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL] `https://arxiv.org/abs/2203.11171`

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] `https://arxiv.org/abs/2201.11903`

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476 [cs.LG] `https://arxiv.org/abs/2503.14476`