DAPO: A Case Study with Small Language Models

Zach Gentile and Alex Lavaee

April 15, 2025

Introduction

Recent advances in Large Language Model (LLM) fine-tuning have demonstrated how reinforcement learning techniques can dramatically enhance reasoning capabilities, with methods like DeepSeek's Group Relative Policy Optimization (GRPO) eliminating the need for separate critic models. Building on this foundation, Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) has achieved impressive results on complex reasoning benchmarks with large models through its innovative clip-higher strategy, dynamic sampling mechanism, token-level policy gradient loss, and overlong reward shaping. While these techniques have proven effective for models with tens of billions of parameters, their applicability to smaller, more accessible models remains unexplored. This project aims to implement DAPO for fine-tuning the compact Gemma 4B model, systematically experimenting with modifications to the algorithm's core components—including refined clip-higher strategies, optimized dynamic sampling, adjusted overlong reward shaping, and the introduction of generalized KL-divergence—to determine whether sophisticated reasoning capabilities can be induced in resource-constrained language models without requiring massive computational resources.

Literature Review

Proximal Policy Optimization (PPO) [3]

Proximal Policy Optimization (PPO) is a policy gradient reinforcement learning algorithm designed to improve training stability and sample efficiency relative to previous methods. PPO iteratively samples data from interaction with the environment and optimizes a specially designed surrogate objective function through stochastic gradient ascent. The key innovation in PPO is its clipped surrogate objective, which prevents excessively large policy updates by restricting the ratio of the new policy probabilities to the old policy within a small, predefined range around 1. This approach retains the advantages of Trust Region Policy Optimization (TRPO), such as stable updates and robustness, while being simpler

to implement, requiring only first-order optimization methods. Empirically, PPO achieves better sample efficiency and overall performance than traditional policy gradient methods and TRPO on a range of benchmark tasks, including simulated robotic locomotion and Atari games.

Group Relative Policy Optimization (GRPO) [2]

The DeepSeek paper introduces an innovative approach for enhancing reasoning capabilities in large language models (LLMs) using reinforcement learning (RL), specifically focusing on Group Relative Policy Optimization (GRPO). Unlike conventional RL methods that rely on separate critic models for evaluating policy performance, GRPO reduces computational costs by estimating performance baselines directly from sampled group outputs.

GRPO works by initially sampling multiple outputs from the existing policy model for each input query. It then optimizes the policy by maximizing an objective function that compares the current policy outputs against these sampled outputs. This optimization process uses a clipped policy ratio to stabilize training, closely related to PPO (Proximal Policy Optimization), but notably avoids using an additional large-scale critic model. The reward signals for optimization in GRPO are derived from both accuracy and format correctness, encouraging the model to provide correct answers and adhere to structured response formats.

Compared to previous methods, GRPO demonstrates superior efficiency by substantially reducing training resource requirements and complexity. DeepSeek's experimental results show significant improvements in reasoning tasks, outperforming conventional supervised fine-tuning methods. Models fine-tuned with GRPO exhibit strong reasoning capabilities, achieving benchmark performance comparable to industry-leading models like OpenAI's o1 series. Thus, GRPO represents a highly effective and resource-efficient approach for training advanced reasoning capabilities into LLMs.

Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) [4]

The Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO) algorithm represents an advancement in large-scale reinforcement learning for large language models (LLMs), explicitly designed to improve reasoning capabilities. Building on the foundations of Group Relative Policy Optimization (GRPO), DAPO addresses several critical issues observed in GRPO implementations, including entropy collapse, reward noise, and training instability.

DAPO differentiates itself from GRPO through four key innovations:

Clip-Higher Strategy: DAPO decouples the clipping thresholds for importance sampling

ratios into separate lower and higher bounds, enhancing the exploration capabilities of the policy. This adjustment prevents entropy collapse, allowing the model to maintain diversity in generated responses.

Dynamic Sampling: DAPO introduces a mechanism to filter out prompts whose outputs provide zero-gradient updates—typically prompts where all sampled outputs are either entirely correct or incorrect. This ensures each batch provides effective gradient updates, improving both training stability and efficiency.

Token-Level Policy Gradient Loss: Unlike GRPO, which averages losses at the sample level potentially biasing against longer responses, DAPO computes losses at the token level. This shift better penalizes undesirable patterns and rewards meaningful, reasoning-rich responses irrespective of their length.

Overlong Reward Shaping: This technique addresses reward noise by introducing soft penalties for excessively long responses, thus promoting conciseness and stability in training.

These improvements yield substantial performance gains. Experimentally validated on the AIME 2024 dataset with the Qwen2.5-32B base model, DAPO significantly outperforms GRPO, achieving a 50 percent accuracy score—surpassing the state-of-the-art 47 percent accuracy of DeepSeek-R1-Zero-Qwen-32B while requiring only half the training steps. The open-source nature of the DAPO algorithm, including code and curated datasets, further underscores its potential for reproducibility and widespread adoption in the community.

Generalized Kullback–Leibler (GKL) Divergence [1]

Generalized Kullback–Leibler (GKL) Divergence Loss explores the limitations of the widely used Kullback–Leibler (KL) Divergence loss by mathematically proving its equivalence to a Decoupled Kullback–Leibler (DKL) loss, which consists of a weighted Mean Square Error (wMSE) component and a Cross-Entropy loss with soft labels. The authors identify two primary issues with the standard KL loss: (1) asymmetric gradient optimization, which neglects the wMSE component in scenarios such as knowledge distillation; and (2) sample-wise biases in predictions that undermine training stability.

To address these problems, the authors propose the Generalized KL (GKL) Divergence Loss. The key innovations in GKL include explicitly breaking the asymmetric gradient optimization to ensure both components (wMSE and Cross-Entropy) contribute effectively to training, and integrating class-wise global information through a smoother weighting function. Specifically, the smoother weight function mitigates convergence challenges by stabilizing training for classes with high predicted scores, thereby enhancing the robustness of training optimization.

Experiments conducted on CIFAR-10/100, ImageNet, and vision-language datasets demonstrate that GKL significantly improves adversarial robustness and knowledge distillation performance, achieving state-of-the-art results on benchmarks such as RobustBench. These results underline GKL's potential as a general-purpose loss function in deep learning scenarios, notably improving both performance and training stability across various tasks.

Proposed RL Techniques

The project employs a systematic approach to investigate DAPO-based fine-tuning of the Gemma 4B model using the AIME dataset. Specifically, the research hypothesizes the following improvements to the DAPO algorithm:

- Replace the dynamic sampling procedure, filtering out prompts with accuracy equal to 1 and 0, with a dynamic sampling procedure that injects adversarial prompts to drive these prompts as artificial examples to improve the model's robustness.
- Replace the soft overlong punishment with a soft perplexity penalty.
- Re-add the KL-divergence term, but use the Generalized KL (GKL) Divergence Loss.

Current Results

We have set up the DAPO environment for inference and training. We have started to establish baselines for the Gemma 4B model on the AIME dataset.

Additionally, we attempted to perform distributed inference on 4 GPUs using the Qwen2.5-32B DAPO model, and were able to partially successful run the inference script (able to generate responses for some sample questions). Though, the results were not as good as we hoped for with the model correctly answering 1/3 of the questions. This might be due to the model's reasoning traces being truncated due to the GPU memory constraints (it was recommended to use 8 GPUs for inference).

Challenges

Several challenges are anticipated in this project:

- **Model capacity limitations**: Small language models have inherently limited capacity, which may constrain their ability to learn complex reasoning patterns. We will need to carefully balance the complexity of reasoning tasks with model capacity.
- **Reward design complexity**: Designing effective reward functions for reasoning tasks is non-trivial. Rewards must be informative enough to guide learning but simple enough to compute efficiently.

- **Computational efficiency**: While SLMs require less computational resources than LLMs, the GRPO training process still demands significant computing power, especially for generating multiple responses per prompt. Optimizing this process will be critical.
- **Overfitting to reward structure**: RL-trained models can exploit patterns in reward functions rather than learning genuine reasoning capabilities. Ensuring generalization beyond the specific training rewards will be essential.
- **Evaluation challenges**: Assessing reasoning capabilities fairly across different model scales requires careful benchmark selection and potentially new evaluation methodologies.

Up until this point we faced quite a few technical issues, including reproducing the package dependencies for the DAPO algorithm inference and training. Specifically, there were issues with missing shared object files in the CUDA library which we resolved. Another struggle was running inference on the Qwen2.5-32B base model which requires 64GB of memory in float16/bfloat16 precision.

Next Steps

We are finishing establishing baselines for the Gemma 4B model on the AIME dataset (base Gemma 4B model compared to DAPO Gemma 4B model). Next, we will be experimenting with modifications and other ablation studies to the DAPO algorithm and observing the effects on performance. Finally we plan on creating a web demo on HuggingFace to showcase the performance of the fine-tuned model.

References

[1] Jiequan Cui et al. *Generalized Kullback-Leibler Divergence Loss*. 2025. arXiv: 2503.08038 [cs.LG]. URL: https://arxiv.org/abs/2503.08038.

[2] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.

[3] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG]. URL: https://arxiv.org/abs/1707.06347.

[4] Qiying Yu et al. *DAPO: An Open-Source LLM Reinforcement Learning System at Scale*. 2025. arXiv: 2503.14476 [cs.LG]. URL: https://arxiv.org/abs/2503.14476.