Verifiable Rewards: Enforcing Correct Reasoning in LLMs via Reinforcement Learning

AMS 691: Topics in Applied Mathematics - Large Language Models

Presented by Group 2: Abhilash, Ananya Sadana, Sumedh Ghavat

Motivation

- LLMs ace “answers” but botch the path

  Shortcut solutions, brittle reasoning, hallucinated steps

- Binary rewards ≠ real understanding

  A 0/1 signal can’t tell why something works

- Verifiable reasoning is now possible

  Math → numeric ground-truth; big LLMs can grade chains-of-thought

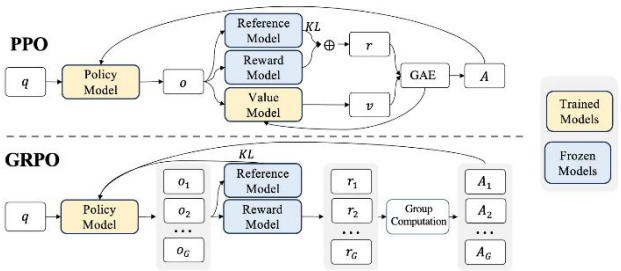
Goal: teach small models how to think, not just what to say by rewarding the process as well as the product

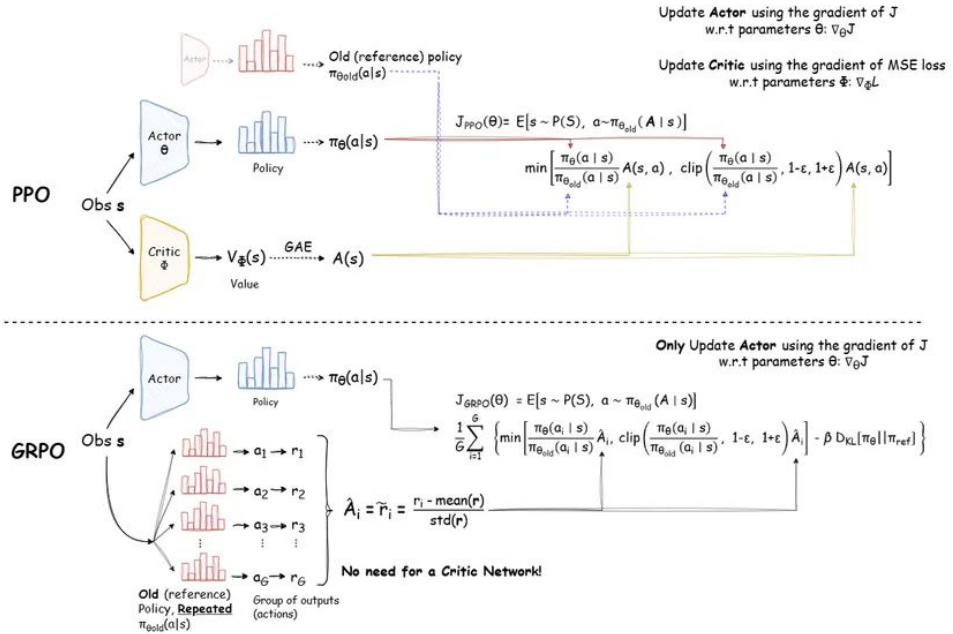
Introduction to GRPO (Group Relative Policy Optimization)

Optimizing policy updates based on groupwise comparisons of sampled actions

- Group based Reward Normalization
- Simplified Objective Function
- Enhanced Stability and Efficiency

GRPO

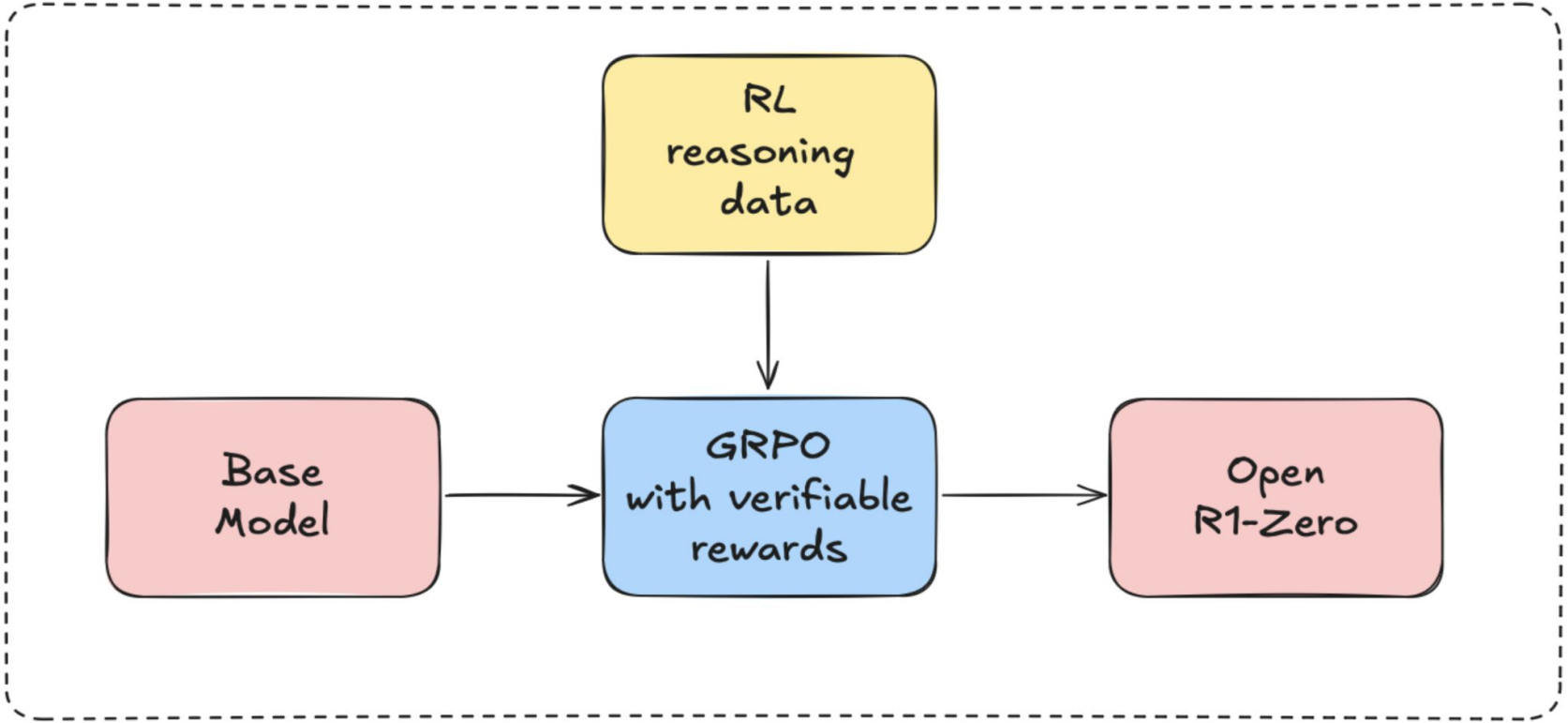

PPO
Policy Model
Reference Model
Reward Model
Value Model
KL
GAE
GRPO
Policy Model
Reference Model
Reward Model
KL
Group Computation
Trained Models
Frozen Models


Update Actor using the gradient of J w.r.t parameters θ: ∇θJ
Update Critic using the gradient of MSE loss w.r.t parameters Φ: ∇ΦL
Old (reference) policy
Actor
Policy
Critic
Value
Obs s
GAE
PPO
GRPO
Only Update Actor using the gradient of J w.r.t parameters θ: ∇θJ
No need for a Critic Network!
Old (reference) Policy, Repeated
Group of outputs (actions)

Introduction to RLVR

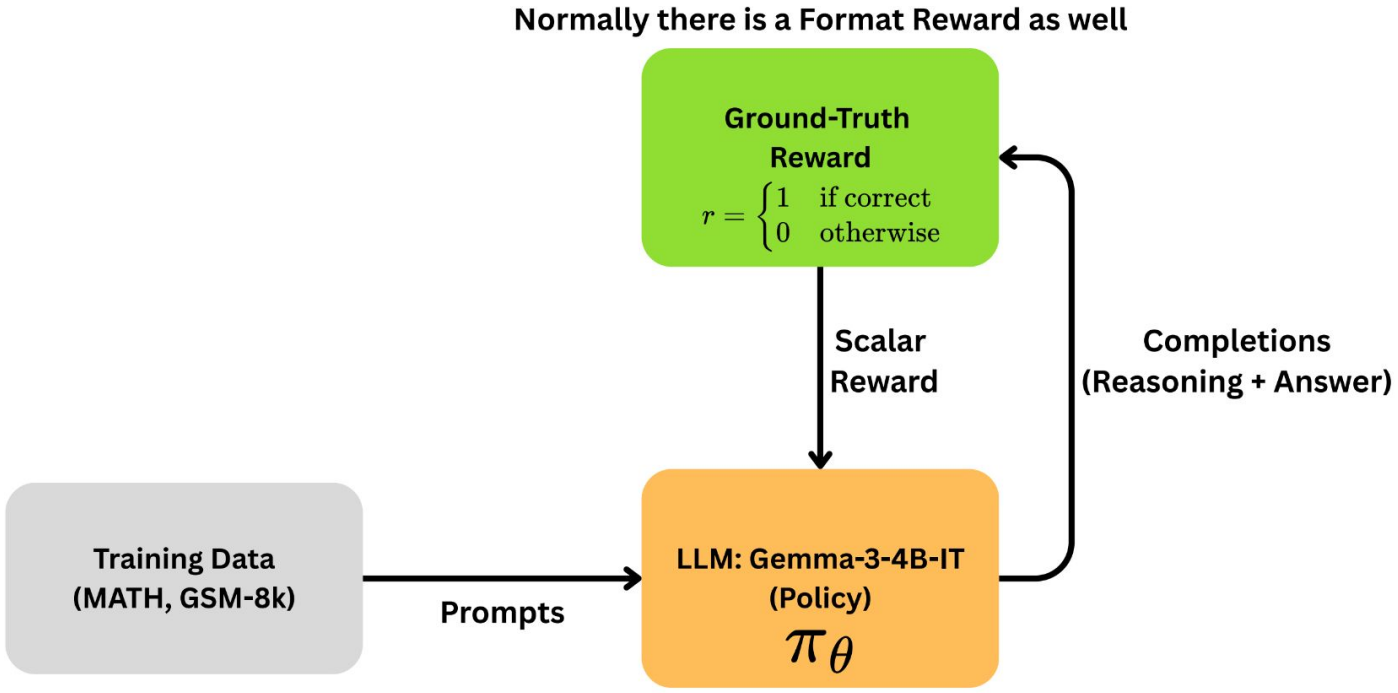
Reinforcement Learning with Verifiable Rewards is a technique where a reinforcement learning model is trained using verifiable rewards instead of relying solely on reward models or human-provided labels. This approach leverages the fact that in certain tasks, like mathematics or coding, there are objective verifiers that can determine if a solution is correct.

Making LLMs Learn How to Think via GRPO Verifiable Rewards Reinforcement Learning With Verifiable Rewards (RLVR)



Baseline Model

Similar to what DeepSeek Math/R1 did at very large scales



A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Baseline Reward Criteria

Format Reward

Enforcing the output to be in a format:

<think>Reasoning</think> <answer>Answer Here</answer>

$$r = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise} \end{cases}$$

Accuracy Reward

Evaluates whether the response is correct inside the answer tags:

$$r = \begin{cases} 1 & \text{if correct} \\ 0 & \text{otherwise} \end{cases}$$

Our Methodology

Apart from the Format and Accuracy Rewards, we added a Reward for correct reasoning with the help of a larger teacher LLM acting as a judge of the students models reasoning capabilities and correctness.

Method 1 (Computationally Cheaper)

Reasoning Correctness Reward

A Larger and more capable LLM evaluates whether the reasoning is correct and up to what extent (LLM calls are only made for the outputs which have been verified to have a correct answer):

$$r = \begin{cases} 1 & \text{Reasoning is fully correct} \\ 0.5 & \text{Reasoning contains some valid steps} \\ 0 & \text{Reasoning is mostly wrong} \end{cases}$$

Method 2 (Computationally Expensive)

Reasoning Correctness Reward

A Larger and more capable LLM evaluates whether the reasoning is correct and up to what extent (LLM calls are always made to reward partial correct reasoning even when the final answer was wrong):

$$r = \begin{cases} 1 & \text{Final answer is correct and reasoning is complete} \\ 0.66 & \text{Final answer is correct, yet reasoning has notable gaps} \\ 0.33 & \text{Final answer is wrong, but reasoning shows some correct steps} \\ 0 & \text{Final answer is wrong and reasoning is largely wrong} \end{cases}$$

Experimental Details

Base/Student LLM: Gemma-3-4B-IT (OpenSource model by Google Deepmind, available on HuggingFace)
Teacher LLM: DeepSeek-R1-Distill-Llama-70B-free (Freely available via Together AI API)

Dataset:

1. **GSM8K (Grade School Math 8K):** is a dataset of 8.5K high quality linguistically diverse grade school math word problems. The dataset was created to support the task of question answering on basic mathematical problems that require multi-step reasoning.

Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?	Natalia sold 48/2 = <<48/2=24>>24 clips in May. Natalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May. #### 72

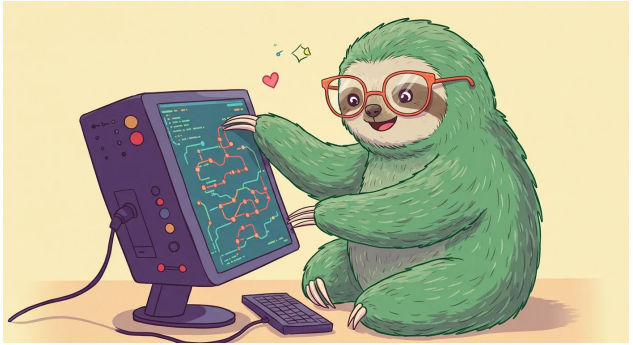
2. **MATH Dataset:** The MATH dataset is a collection of mathematics competition problems designed to evaluate mathematical reasoning and problem-solving capabilities in computational systems. Containing 12,500 high school competition-level mathematics problems, this dataset is notable for including detailed step-by-step solutions alongside each problem.

How many vertical asymptotes does the graph of $y=\frac{2}{x^2+x-6}$ have?	The denominator of the rational function factors into $x^2+x-6=(x-2)(x+3)$. Since the numerator is always nonzero, there is a vertical asymptote whenever the denominator is 0, which occurs for $x = 2$ and $x = -3$. Therefore, the graph has $\boxed{2}$ vertical asymptotes.

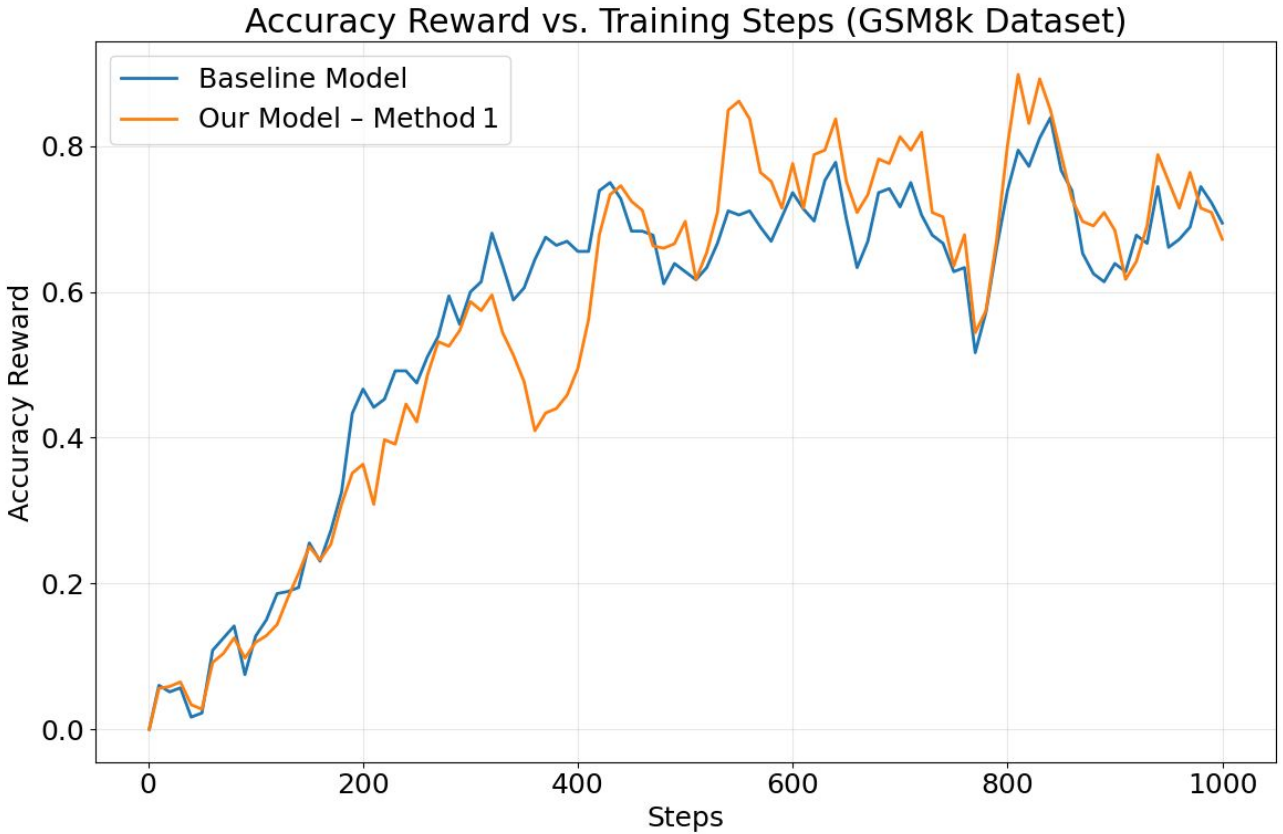
Experimental Details

Unsloth — Turbo-charged QLoRA

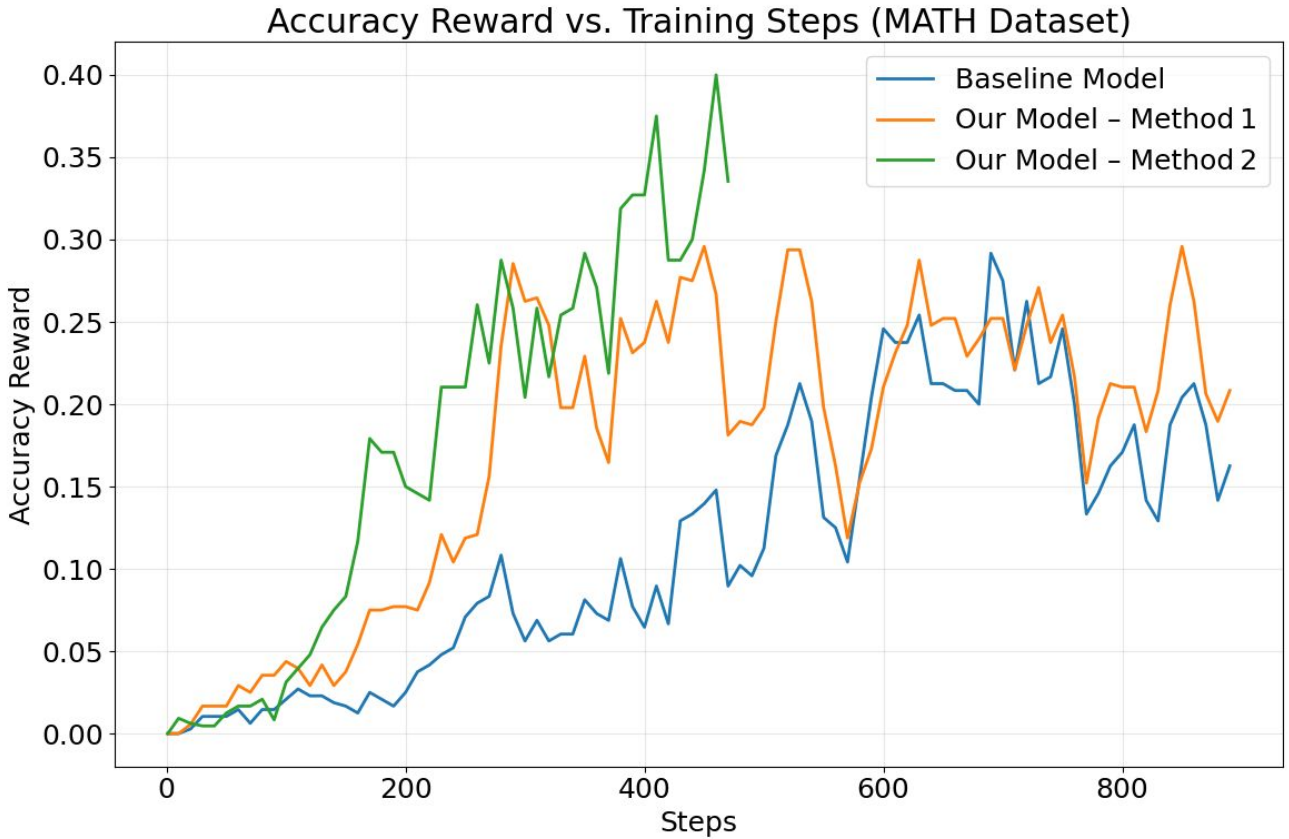
- Drop-in wrapper around HuggingFace models
- Combines **LoRA** (rank 32) plus **4-bit weight quantization** (**QLoRA**)
- Custom Triton kernels & FlashAttention 2 ⇒ up to **2.7x faster, 74 % less RAM**
- Gemma-4B + adapters fits on a single **Colab A100/L4** GPU



Results (Plots for GSM8K)



Results (Plots for MATH)



Results

Accuracy of the models trained and the base model (Gemma-3-4B-IT) on the test sets of the datasets. A simple exact matching is used for GSM8k and a Sympy based matching for MATH-500

	MATH-500@1	**GSM-8K@1**
Gemma-3-4B-IT	10.71%	25.57%
Baseline Model (RLVR)	27.32%	78.92%
Our Model - Method 1 (RLVR)	33.11%	81.31%
Our Model - Method 2 (RLVR)	33.05%	-

Conclusion

Gap: Binary-reward RL fixes answers but not reasoning.

Proposed Solution: RL with Verifiable Rewards

- format ✔ numeric answer ✔ LLM-judge-graded reasoning ✔
- Finetuned Gemma-3-4B-IT via GRPO

Gains:

- GSM-8K 78.92 → 81.3 % (**+2 pp** over binary RL)
- MATH-500 27.32 → 33.1 % (**+6 pp** over binary RL)

QnA