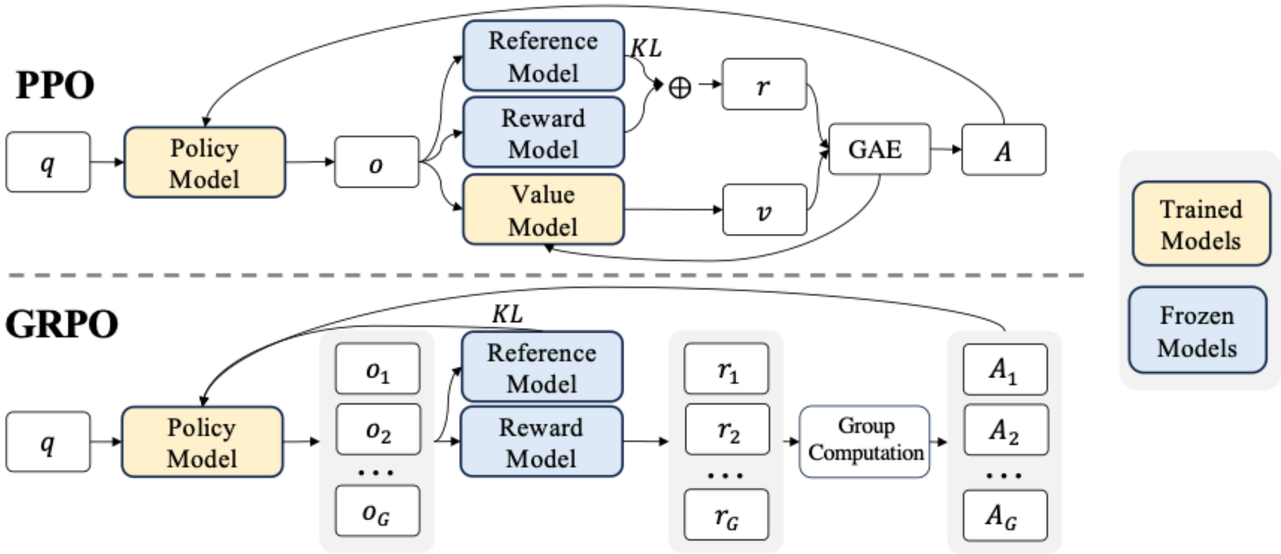
Group Relative Policy Optimization (GRPO)

- GRPO generates multiple trajectories per question, and uses the averaged (normalized) reward as advantage. —> no need for a value model!



DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao[1,2*†], Peiyi Wang[1,3*†], Qihao Zhu[1,3*†], Runxin Xu[1], Junxiao Song[1]
Xiao Bi[1], Haowei Zhang[1], Mingchuan Zhang[1], Y.K. Li[1], Y. Wu[1], Daya Guo[1*]

$$A_i = \frac{r_i - mean(r)}{std(r)}$$

GRPO is now commonly used for training reasoning models.

Exponentially rare rewards

- What if the questions are too hard for the model?
  - E.g., if the chance of generating a correct solution is low and rewards are sparse?
  - E.g., if a solution has n steps and chance of doing each step correctly is 0.5, the chance of having a correct solution would be 2^{-n}.

$$A_i = \frac{r_i - mean(r)}{std(r)}$$

Chain of parities

A contextual blind cliff walk [Schaul et al. ’15]

input: $x_1, x_2, \ldots, x_n$

output: $y_1, z_1, y_2, z_2, \ldots, y_n, z_n$

x_i : -1 or 1 (uniform)

y_i : uncostrained, can be -1 or 1

$z_i = z_{i-1} \times y_i \times x_i$ i.e., $(x_1, \ldots, x_i) \rightarrow z_i = x_1 \times \ldots \times x_i \times y_1 \times \ldots \times y_i$

- An environment where a solution has n steps.
- Each question has 2^n possible solutions.
- The chance of finding a correct solution randomly would be 2^{-n}.

Chain of parities

A contextual blind cliff walk [Schaul et al. ’15]

input: $x_1, x_2, \ldots, x_n$

output: $y_1, z_1, y_2, z_2, \ldots, y_n, z_n$

$$z_i = z_{i-1} \times y_i \times x_i \quad \text{i.e., } (x_1, \ldots, x_i) \to z_i = x_1 \times \ldots \times x_i \times y_1 \times \ldots \times y_i$$

Assume access to a small dataset $\{(\bar{x}^t, \bar{y}^t, \bar{z}^t)\}_{t=1}^{D}$

- RL training fails: discovering a single valid output via random exploration becomes exponentially unlikely.
- Supervised training fails at low k: the task involves learning n – 1 parity functions of degree three.
- SFT + RL also fails: Weak learning in parity happens together with strong learning

Issues with standard practice—solution

- What if the questions are too hard for the model?
  - E.g., if the chance of generating a correct solution is low and rewards are sparse?
  - E.g., if a solution has n steps and chance of doing each step correctly is 0.5, the chance of having a correct solution would be 2^{-n}.
- We can help the model by revealing the first part of the solution and asking the model to complete it.
  - Conditioning on the first $n - k$ steps boosts the chance of reaching a solution to 2^{-k}.
  - During training, we'd like to reduce the amount of the solution that is given.

Adaptive Backtracking (AdaBack)[1]

Guiding the model by adaptively revealing rationales

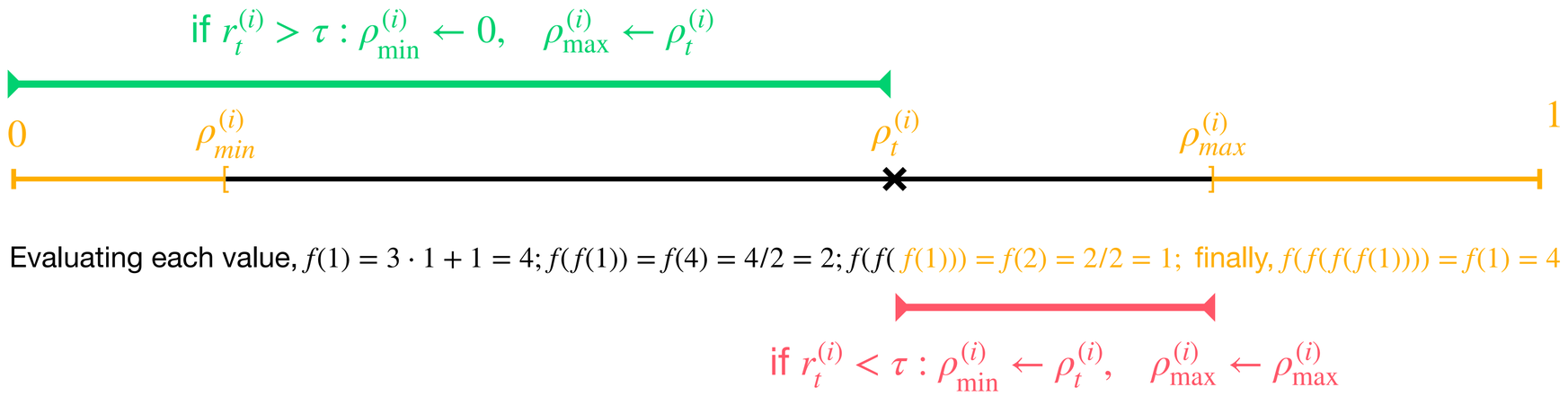
Question: Let $f(x) = \begin{cases} x/2 & \text{if } x \text{ is even,} \\ 3x+1 & \text{if } x \text{ is odd .} \end{cases}$ what is $f(f(f(f(1))))$?

Ground-truth ans.: Evaluating each value, $f(1) = 3 \cdot 1 + 1 = 4; f(f(1)) = f(4) = 4/2 = 2; f(f(f(1))) = f(2) = 2/2 = 1;$ finally, $f(f(f(f(1)))) = f(1) = 4$

$\rho_t^{(i)}$: sampled portion for question i at epoch t

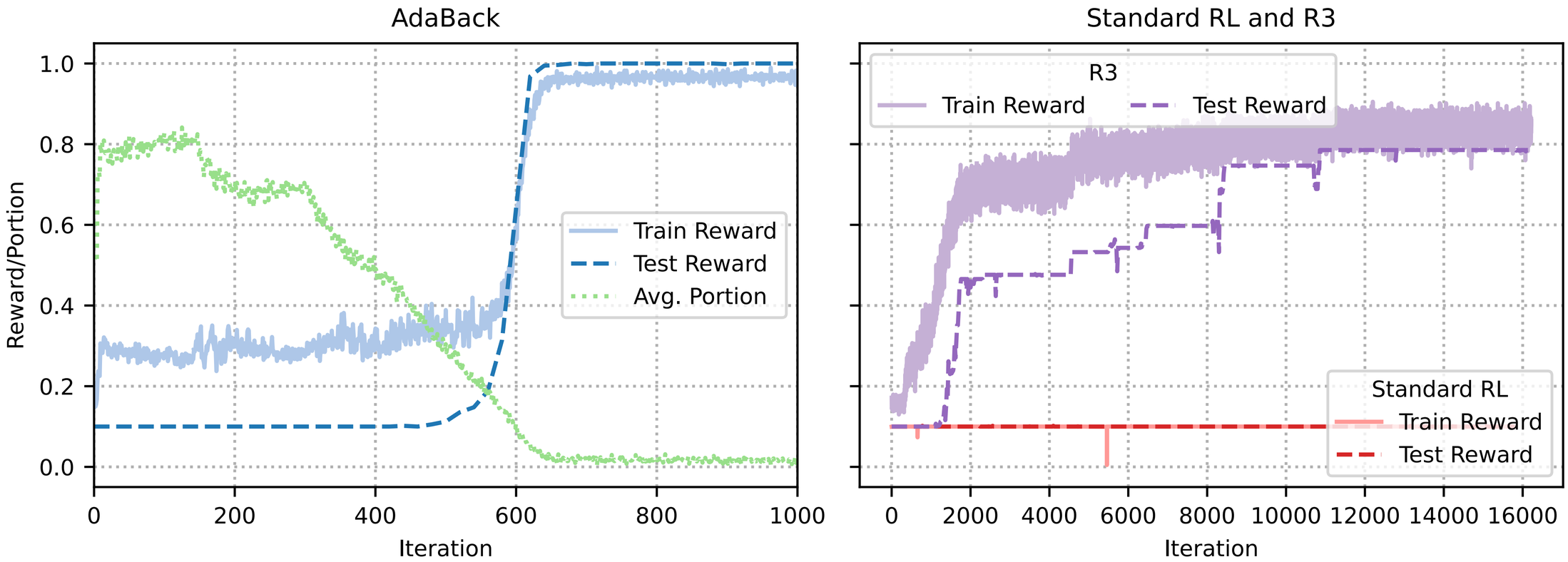
r: the average reward for this question given by GRPO

depending on r (difficulty of the sample), we reveal more or less at next epoch



[1] RL for Reasoning by Adaptively Revealing Rationales

Separation result on chain-of-parities dataset



[R3: Z. Xi, W. Chen, et. al. Training large language models for reasoning through reverse curriculum reinforcement learning, ICML 2024]

Results

GSM8k questions in base 7

Concatenation of two GSM8k questions

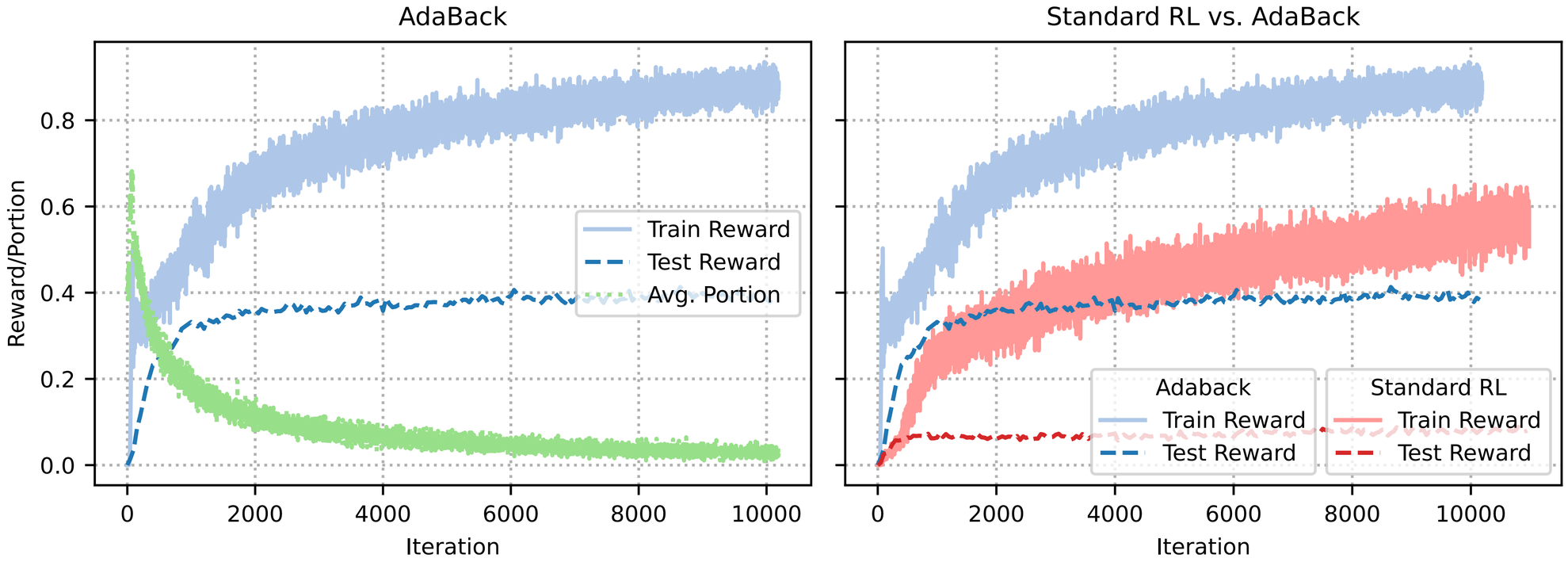
Table 1: Final test accuracy for each method across tasks and model sizes.

Method	MATH		GSM8k		Base-7 GSM8k		Tensor-2 GSM8k	
	1B	3B	1B	3B	1B	3B	1B	3B
Base+RL	6.4	15.0	7.9	63.7	4.8	4.9	0.0	0.0
SFT+RL	7.4	17.7	36.7	72.7	14.4	45.4	6.9	42.7
AdaBack	9.1	19.1	39.2	**73.3**	18.4	43.9	8.5	**49.2**
SFT+AdaBack	**9.5**	**19.9**	**43.2**	70.7	**24.5**	**49.9**	**11.3**	42.2

- Particularly advantageous when tasks are hard for models:
  - Small models
  - Models without SFT
  - New tasks

RL training on GSM8k dataset

Llama3 1B without supervised fine-tuning



RL training for Base-7 GSM8k

Llama3 1B SFTed on Base-7 GSM8k

GSM8k questions in base 7