GPO: Learning from Critical Steps to Improve LLM Reasoning

Author: Jiahao Yu, Zelei Cheng, Xian Wu, Xinyu Xing
Presenter: Jiahao Yu

Northwestern

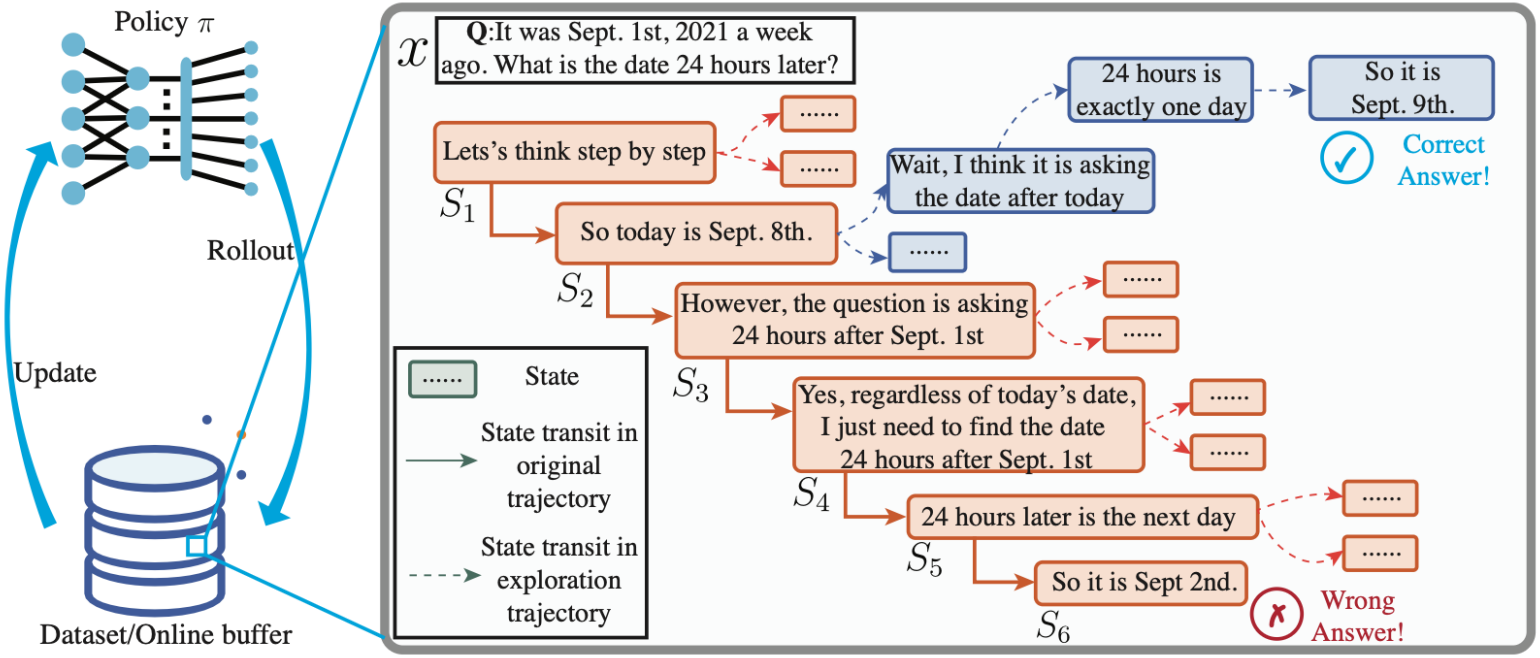
Meta

CapitalOne

Motivation

1. LLM Reasoning can be brittle
2. Standard fine-tuning treat reasoning trajectories as a whole
3. Previous method fail to pinpoint which specific step was critical

Our Solution: GPO



A General Framework

Algorithm 1: GPO Optimization Framework

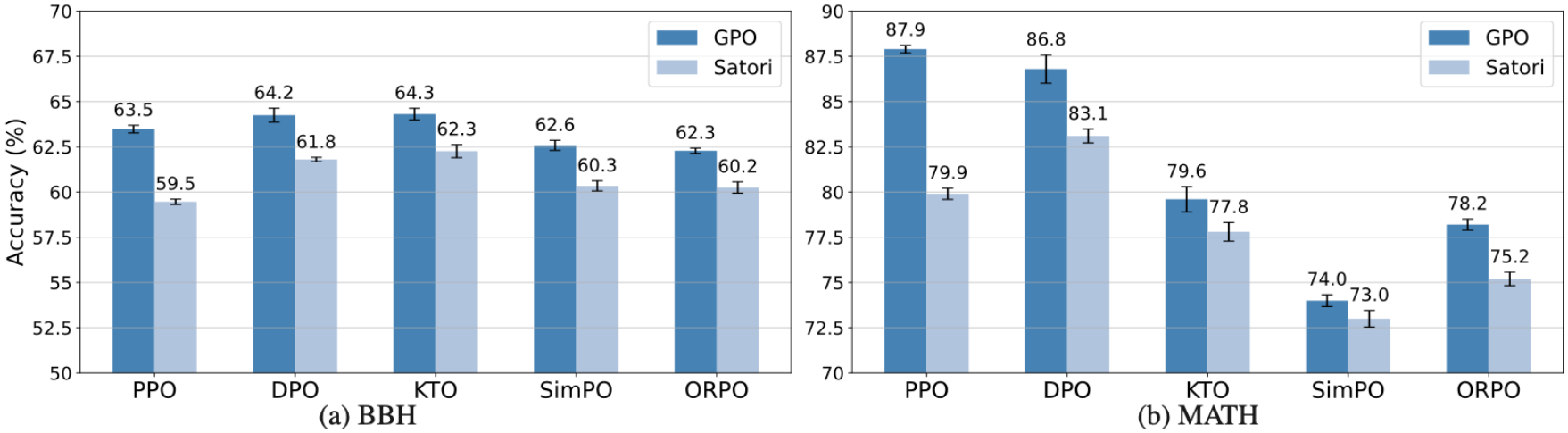
Procedure-I: Online Policy Training (PPO-based)
Input: Initial LLM policy $\pi^0 = \pi_{\text{ref}}$, reasoning dataset D_r
for iteration = 1, 2, ..., T **do**
  $\mathcal{D} \leftarrow \emptyset$
  for n = 1, 2, ... **do**
    Random sampling a question x from the reasoning dataset D_r
    Run π^t to generate a K-step reasoning trajectory $y = (y_0, \ldots, y_{K-1})$ split by newlines
    Identify the critical step y_m with maximal advantage $A^{\pi^t}(x, y_{0:i-1}; y_i)$
    Reset π^t to y_m and roll-out π^t to generate trajectory $y' = (y_m, \ldots, y'_{K-1})$
    Add trajectory y' and the final reward r to $\mathcal{D}$
  end for
  Optimize π^t with respect to the policy gradient loss (*e.g.*, PPO loss) in Eqn. 1 on $\mathcal{D}$
end for

Procedure-II: Preference Data Generation and Optimization (DPO-based)
Input: Supervised-finetuned base policy π_{ref}, reasoning dataset D_r
$\mathcal{D} \leftarrow \emptyset$
for iteration = 1, 2, ..., T **do**
  Repeat the sampling, trajectory generation using π_{ref}, and critical step identification as in **Procedure-I** to extract the important step y_m from trajectory y.
  Generate two continuations starting from y_m to obtain a positive trajectory $y^+ = (y_0, \ldots, y_m, \ldots, y^+_{K-1})$ and a negative trajectory $y^- = (y_0, \ldots, y_m, \ldots, y^-_{K-1})$
  Add the preference pair (x, y^+, y^-) to $\mathcal{D}$
end for
Optimize π with respect to the preference loss (*e.g.*, DPO loss) in Eqn. 2 on $\mathcal{D}$

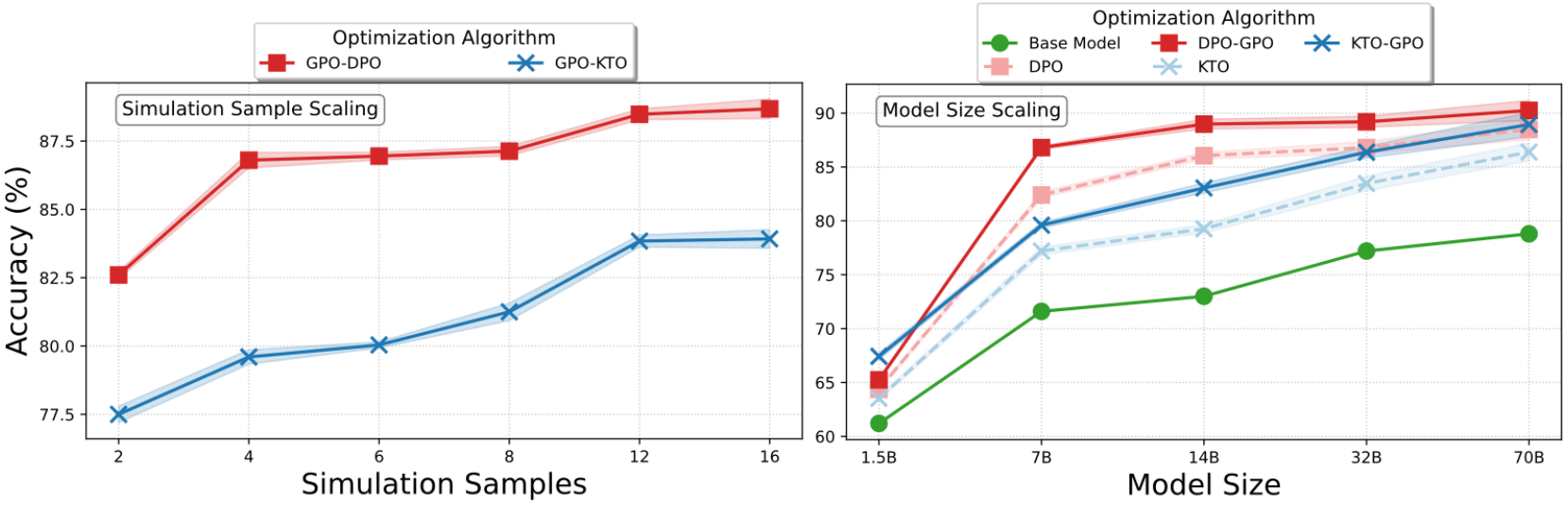
Experimental Results

Algorithms	Test Accuracy (%)						
	BBH	MATH	GSM8K	MMLU	MMLUPro	AIME-2024	AIME-2025
Base Model	59.97	71.60	86.50	54.09	38.80	13.33	16.67
PPO	61.82	79.60	86.96	56.66	47.47	26.67	23.33
GPO-PPO	**63.48**	**87.80**	**87.44**	**59.39**	**51.05**	**30.00**	**26.67**
DPO	63.20	82.40	86.05	57.08	48.28	20.00	20.00
GPO-DPO	**64.25**	**86.80**	**88.48**	**58.93**	**51.93**	**26.67**	**26.67**
KTO	62.86	77.20	89.31	59.42	49.02	20.00	20.00
GPO-KTO	**64.31**	**79.60**	**90.25**	**61.35**	**50.52**	**23.33**	**26.67**
SimPO	61.97	72.20	86.58	56.93	45.70	20.00	23.33
GPO-SimPO	**62.58**	**74.00**	**88.35**	**57.44**	**47.74**	**23.33**	**26.67**
ORPO	61.75	75.20	87.26	57.72	46.66	20.00	20.00
GPO-ORPO	**62.28**	**78.20**	**88.17**	**58.72**	**48.65**	**23.33**	**23.33**

Ablation Study



Scaling Behavior



Human Alignment

A user study found a strong correlation. Steps GPO identified as 'critical' were also chosen by human evaluators **44-88%** of the time as the most pivotal moment in the reasoning process.

Thank you!

Code: https://github.com/sherdencooper/GPO