中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Teaching Large Language Models to Reason with Reinforcement Learning

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy

Meta
Georgia Institute of Technology

Background

高瓴人工智能学院
Gaoling School of Artificial Intelligence

- **The reasoning abilities of large language models (LLMs)**
  - Math, science, and code benchmarks…
- **Technique routine**
  - Prompting strategies: CoT, ToT…
  - Fine-tuning: SFT, RL-based fine-tuning

Proximal Policy Optimization (**PPO**)
Expert iteration (**EI**)
Return-Conditioned RL (**RCRL**)

? How to improve the reasoning capabilities across a variety of reward schemes and model initializations.

Proximal Policy Optimization (PPO)

Policy gradient

- **Agent accomplishes the task step by step and interact with environment**
  - A trajectory: $\tau = \{s_1, a_1, s_2, a_2, \cdots, s_t, a_t\}$
  - Agent-p_θ, env.-p

$$p_\theta(\tau) = p(s_1)\, p_\theta(a_1|s_1)\, p(s_2|s_1, a_1)\, p_\theta(a_2|s_2)\, p(s_3|s_2, a_2) \cdots$$
$$= p(s_1) \prod_{t=1}^{T} p_\theta(a_t|s_t)\, p(s_{t+1}|s_t, a_t)$$

- **Maximize reward**
  - Gradient ascent:

$$\bar{R}_\theta = \sum_\tau R(\tau) p_\theta(\tau) = \mathbb{E}_{\tau \sim p_\theta(\tau)}[R(\tau)] \quad \longrightarrow \quad \nabla \bar{R}_\theta = \sum_\tau R(\tau) \nabla p_\theta(\tau)$$

Proximal Policy Optimization (PPO)

Policy gradient

$$\bar{R}_\theta = \sum_\tau R(\tau) p_\theta(\tau) = \mathbb{E}_{\tau \sim p_\theta(\tau)}[R(\tau)] \longrightarrow \nabla \bar{R}_\theta = \sum_\tau R(\tau) \nabla p_\theta(\tau)$$

$$\frac{\nabla p_\theta(\tau)}{p_\theta(\tau)} = \nabla \log p_\theta(\tau) \qquad \boxed{\nabla f(x) = f(x) \nabla \log f(x)}$$

$$\begin{aligned}
\nabla \bar{R}_\theta &= \sum_\tau R(\tau) \nabla p_\theta(\tau) \\
&= \sum_\tau R(\tau) p_\theta(\tau) \frac{\nabla p_\theta(\tau)}{p_\theta(\tau)} \\
&= \sum_\tau R(\tau) p_\theta(\tau) \nabla \log p_\theta(\tau) \\
&= \mathbb{E}_{\tau \sim p_\theta(\tau)}[R(\tau) \nabla \log p_\theta(\tau)] \approx \frac{1}{N} \sum_{n=1}^{N} R\left(\tau^n\right) \nabla \log p_\theta\left(\tau^n\right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} R\left(\tau^n\right) \nabla \log p_\theta\left(a_t^n \mid s_t^n\right)
\end{aligned}$$

$$\theta \leftarrow \theta + \eta \nabla \bar{R}_\theta$$

Proximal Policy Optimization (PPO)

高瓴人工智能学院
Gaoling School of Artificial Intelligence

$$\nabla \bar{R}_\theta = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[R(\tau) \nabla \log p_\theta(\tau) \right]$$

PPO

- **On policy**

$$\nabla \bar{R}_\theta = \mathbb{E}_{\tau \sim p_{\theta'}(\tau)} \left[\frac{p_\theta(\tau)}{p_{\theta'}(\tau)} R(\tau) \nabla \log p_\theta(\tau) \right]$$

- **Advantage**

$$\mathbb{E}_{(s_t, a_t) \sim \pi_\theta} \left[A^\theta(s_t, a_t) \nabla \log p_\theta(a_t^n | s_t^n) \right]$$

$$\mathbb{E}_{(s_t, a_t) \sim \pi_{\theta'}} \left[\frac{p_\theta(s_t, a_t)}{p_{\theta'}(s_t, a_t)} A^\theta(s_t, a_t) \nabla \log p_\theta(a_t^n | s_t^n) \right]$$

Proximal Policy Optimization (PPO)

PPO

$$\mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(s_t,a_t)}{p_{\theta'}(s_t,a_t)}A^{\theta}(s_t,a_t)\nabla\log p_\theta(a_t^n|s_t^n)\right]$$

$$\mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}\cancel{\frac{p_\theta(s_t)}{p_{\theta'}(s_t)}}A^{\theta'}(s_t,a_t)\nabla\log p_\theta(a_t^n|s_t^n)\right]$$

$$\mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}A^{\theta'}(s_t,a_t)\nabla\log p_\theta(a_t^n|s_t^n)\right]$$

$$\boxed{\nabla f(x)=f(x)\nabla\log f(x)}$$

$$J^{\theta'}(\theta)=\mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}A^{\theta'}(s_t,a_t)\right]$$

Proximal Policy Optimization (PPO)

PPO

$$J^{\theta'}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{p_\theta\left(a_t|s_t\right)}{p_{\theta'}\left(a_t|s_t\right)}A^{\theta'}\left(s_t,a_t\right)\right]$$

$$J_{\mathrm{PPO2}}^{\theta^k}(\theta) \approx \sum_{(s_t,a_t)} \min\left(\frac{p_\theta\left(a_t|s_t\right)}{p_{\theta^k}\left(a_t|s_t\right)}A^{\theta^k}\left(s_t,a_t\right),\right.$$
$$\left.\mathrm{clip}\left(\frac{p_\theta\left(a_t|s_t\right)}{p_{\theta^k}\left(a_t|s_t\right)},1-\varepsilon,1+\varepsilon\right)A^{\theta^k}\left(s_t,a_t\right)\right)$$

Expert iteration (EI)

高瓴人工智能学院 Gaoling School of Artificial Intelligence

- **Expert policy approximation $\widehat{\pi}_0^*$**
  - Generate K rollouts and **filter out** incorrect solutions and duplicates to construct D_1
- **Distilled back into a policy π_1**

$$\sum_{\tau \in D} \sum_{t=1}^{H} -\log\left(\pi_\theta\left(a_t \mid s_t\right)\right)$$

- **Repeated fine-tuning**
  - This process can be repeated to construct policy π_i fine-tuned on dataset $D_i = R_i \cup D_{i-1}$ where R_i corresponds to exploration done by π_{i-1}.

Return-Conditioned RL (RCRL)

- **Train policies conditioned on both the current state s and desired return R**

$$\sum_{\tau \in D} \sum_{t=1}^{H} -\log\left(\pi_\theta\left(a_t \mid s_t, g_t\right)\right)$$

- **Data construction**
  - Best **EI** policy sample K many times from $\{s_1, s_2, \ldots, s_i\}$, and $\{l_1, l_2, \ldots, l_K\}$ evaluating the correctness of the generated final answers.
  - s_i is labeled as "[GOOD]" if $\frac{1}{K}\sum_{k=1}^{K} l_k \geq T$.

Experiments

高瓴人工智能学院
Gaoling School of Artificial Intelligence

- **GSM8K [w/ sft]**

From pre-train model

From sft model

	maj@1		maj@96		rerank@96†		pass@96	
	7B	13B	7B	13B	7B	13B	7B	13B
SFT	0.41	0.48	0.47	0.53	0.54	0.68	0.72	0.84
EI_n	**0.48**	**0.53**	**0.55**	**0.59**	0.64	**0.71**	0.8	**0.88**
ORM EI_n	**0.48**	**0.53**	0.54	0.58	**0.65**	**0.71**	**0.81**	0.87
ORM RCRL	0.45	0.51	0.5	0.56	0.54	0.69	0.73	0.83
Sparse PPO	0.44	0.51	0.49	0.55	0.58	0.67	0.77	0.85
Dense PPO	0.43	0.50	0.47	0.54	0.53	0.65	0.71	0.81
Sparse ORM PPO	0.46	0.51	0.51	0.55	0.59	0.67	0.79	0.83
Dense ORM PPO	0.46	0.51	0.52	0.55	0.59	0.67	0.76	0.83
Llema*	0.40	0.62	0.54	0.69	N/A		N/A	
RFT	0.47	0.54	0.58	0.65	N/A		N/A	
WizardMath	0.55	0.64	N/A		N/A		N/A	
GPT-3**	0.2	0.31	N/A		0.39	0.55	0.71	NA
GPT-4***	0.91		N/A		N/A		N/A	

Experiments

➢ Analysis experiment

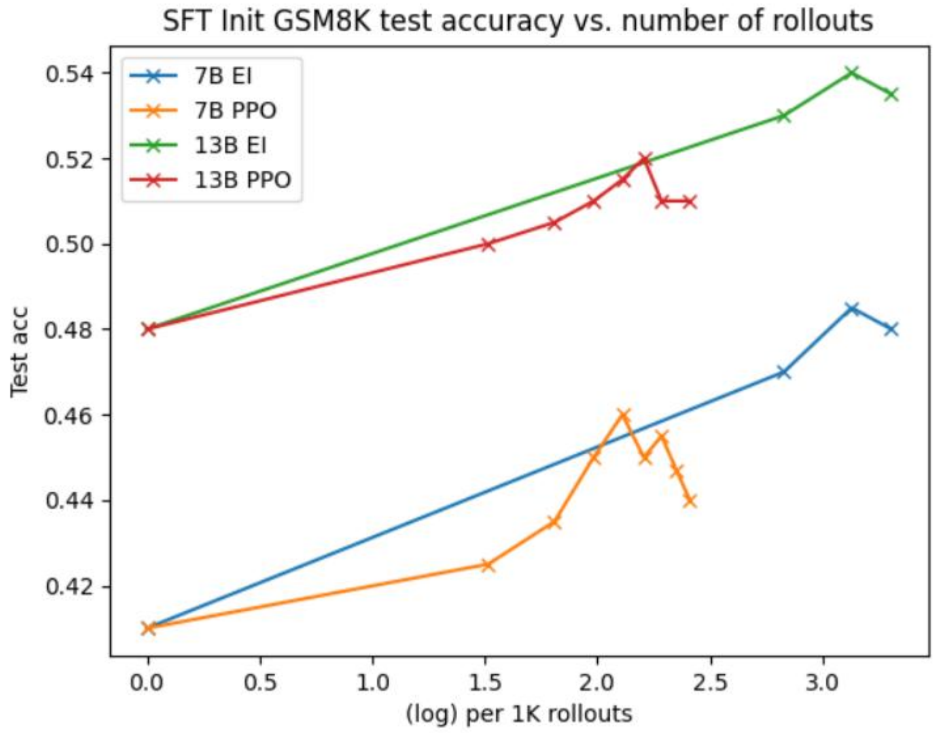

Figure 1 Sample complexities of SFT initialized models on GSM8K. EI achieves better performance than PPO with the same order of magnitude of samples.

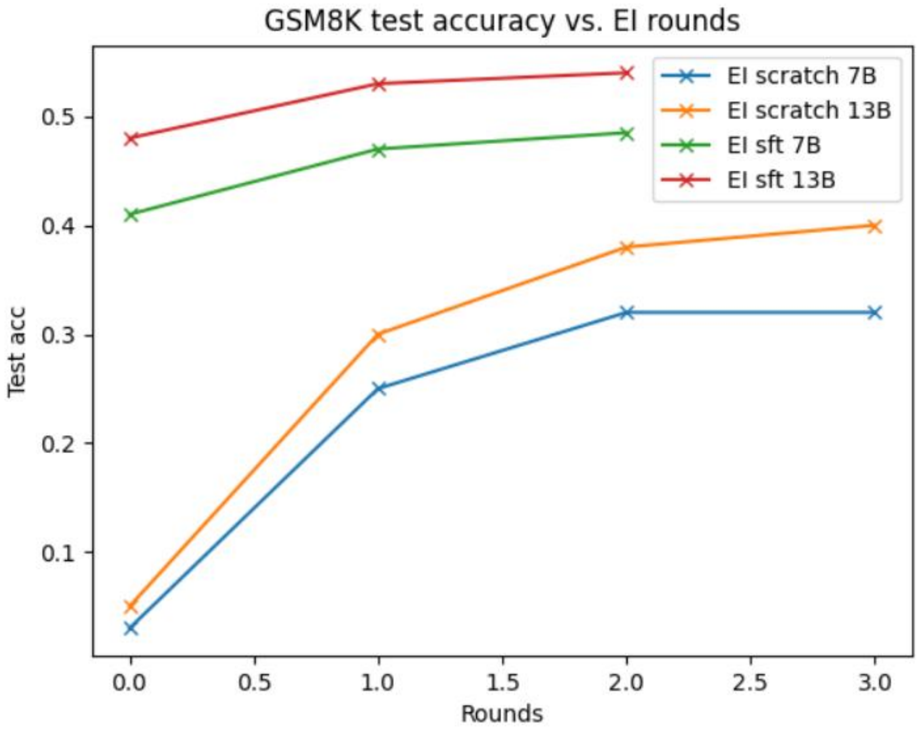

Figure 2 Accuracy of EI models on GSM8K test vs. number of iterations. Performance seems plateaus for SFT initialized models after two iterations. The pretrained checkpoints converge after four iterations.

Experiments

- **GSM8K [w/o sft]**

	maj@1		maj@n		rerank@n†		pass@n	
	7B	13B	7B	13B	7B	13B	7B	13B
Prompted	0.05	0.03	0.14	0.18	0.17	0.24	0.22	0.27
EI_n	0.31	0.4	0.35	0.47	0.39	0.63	0.45	**0.83**
ORM EI	0.28	0.37	0.33	0.43	0.37	0.59	0.42	0.76
Sparse PPO	**0.32**	**0.41**	**0.37**	**0.48**	**0.41**	**0.65**	**0.5**	**0.83**
Sparse ORM PPO	0.29	0.38	0.34	0.44	0.4	0.62	0.49	0.81
Dense ORM PPO	0.29	0.39	0.35	0.45	0.41	0.64	0.5	0.82

Conclusion

- A comprehensive study of PPO fine-tuning of LLMs on reasoning tasks using different types of rewards
- A comparison to expert iteration and return-conditioned RL from which we find expert iteration reliably attains the best performance and competitive sample complexity across the board..

中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Thank You for listening!