Harvard John A. Paulson
School of Engineering
and Applied Sciences

Kempner
INSTITUTE
at Harvard University

The Power of Resets!

Learning better, one reset at a time.

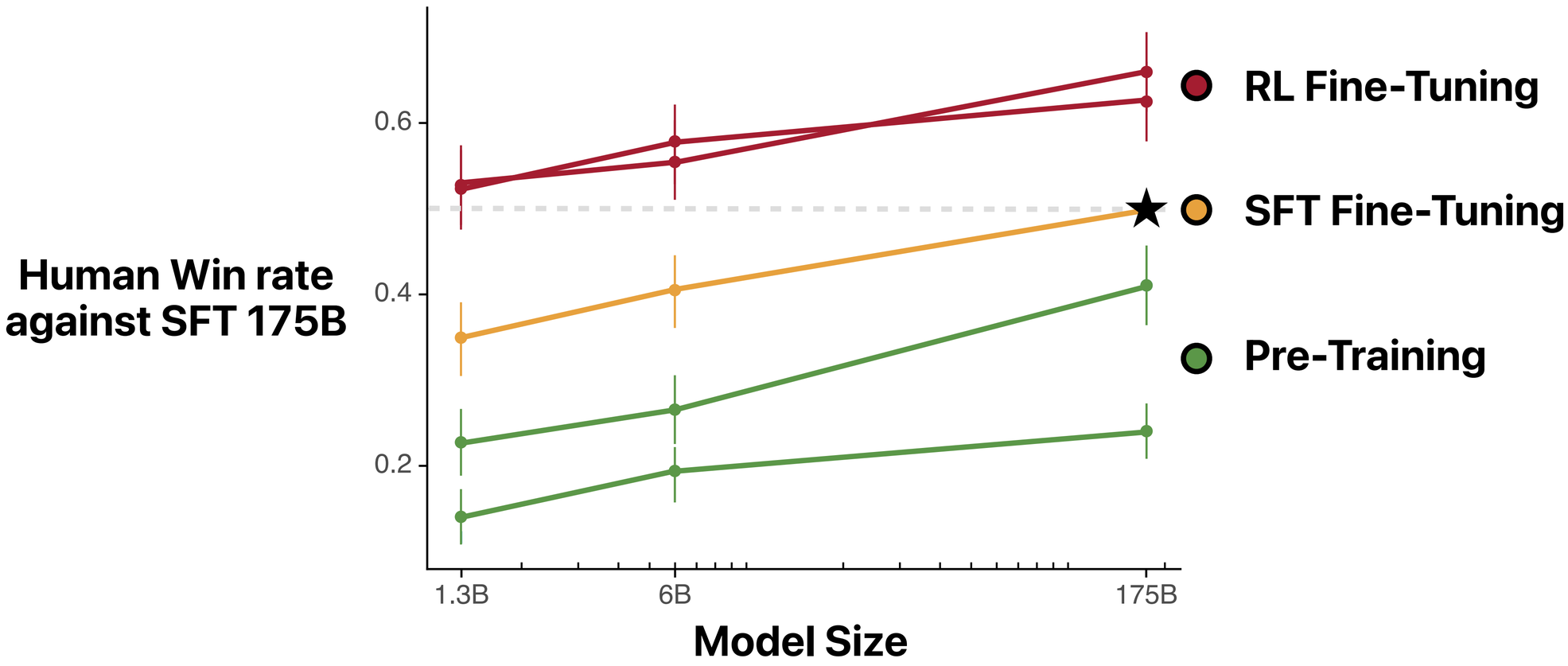
Kianté Brantley, Assistant Professor, Harvard University

"A large language model is a deep learning model trained on massive text datasets to understand and generate human language."

Large Language Model

How do *different* objectives shape LLM behavior?

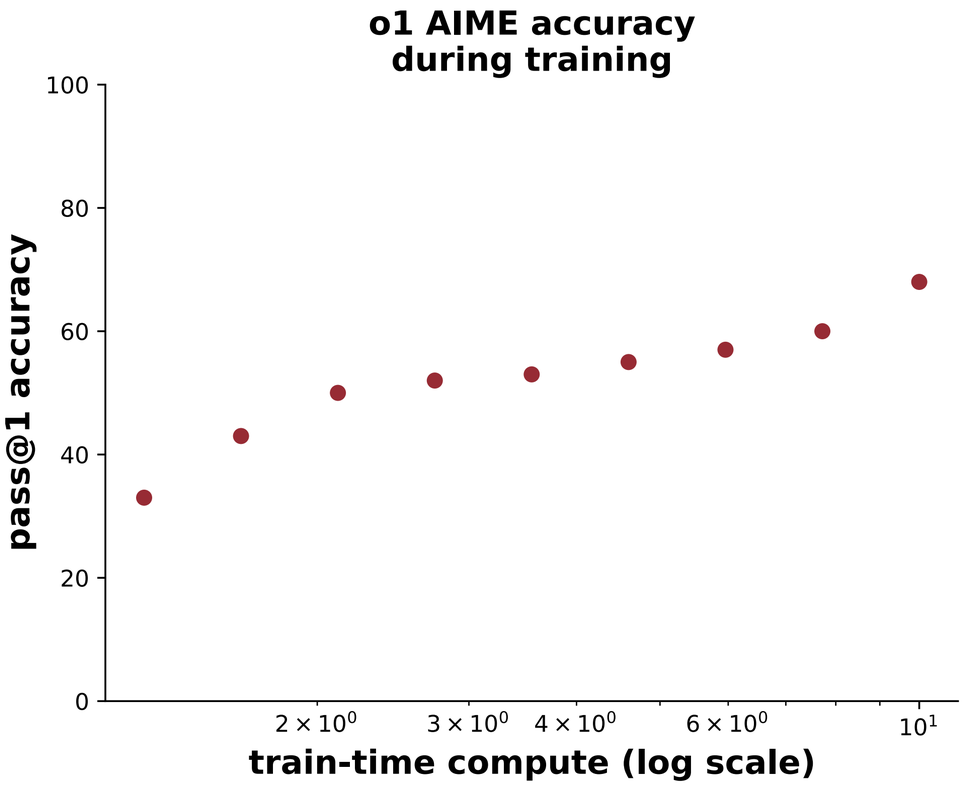
Learning from Human Feedback

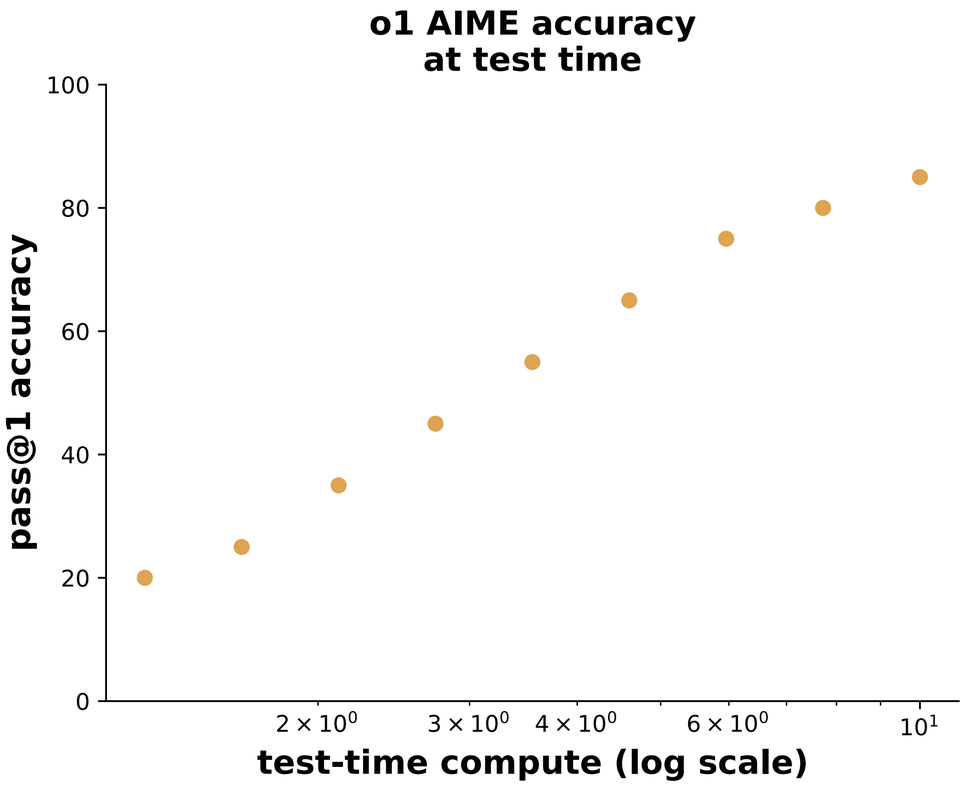


Long Ouyang et al. Training language models to follow instructions with human feedback OpenAI 2022

Learning to reason with LLMs

“We are introducing OpenAI o1, a new large language model trained with reinforcement learning to perform complex reasoning.” - Openai





MinxEnt RL

$$J(\pi) = \mathbb{E}_{\pi}\underbrace{\left[r(x, y)\right]} - \frac{1}{\eta}\underbrace{\mathbb{D}_{\mathrm{KL}}(\pi \,\|\, \pi_{\mathrm{ref}})}$$

Maximize reward **under** **Minimal deviation**

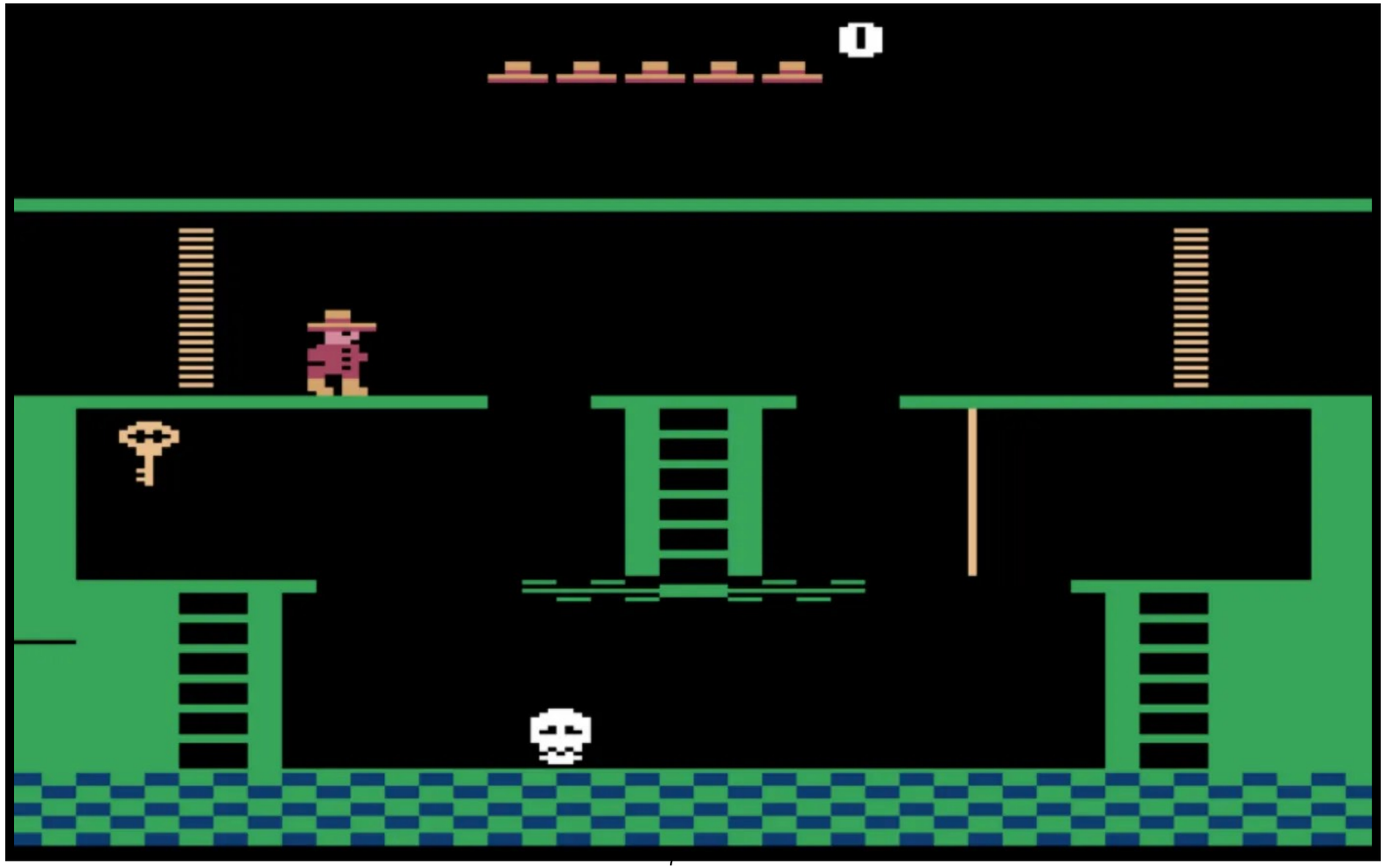
MinxEnt RL

When you say, "**This is a reinforcement learning problem**," you should say it with the same excitement as "**This is NP-hard**." **- Tim Vieira**

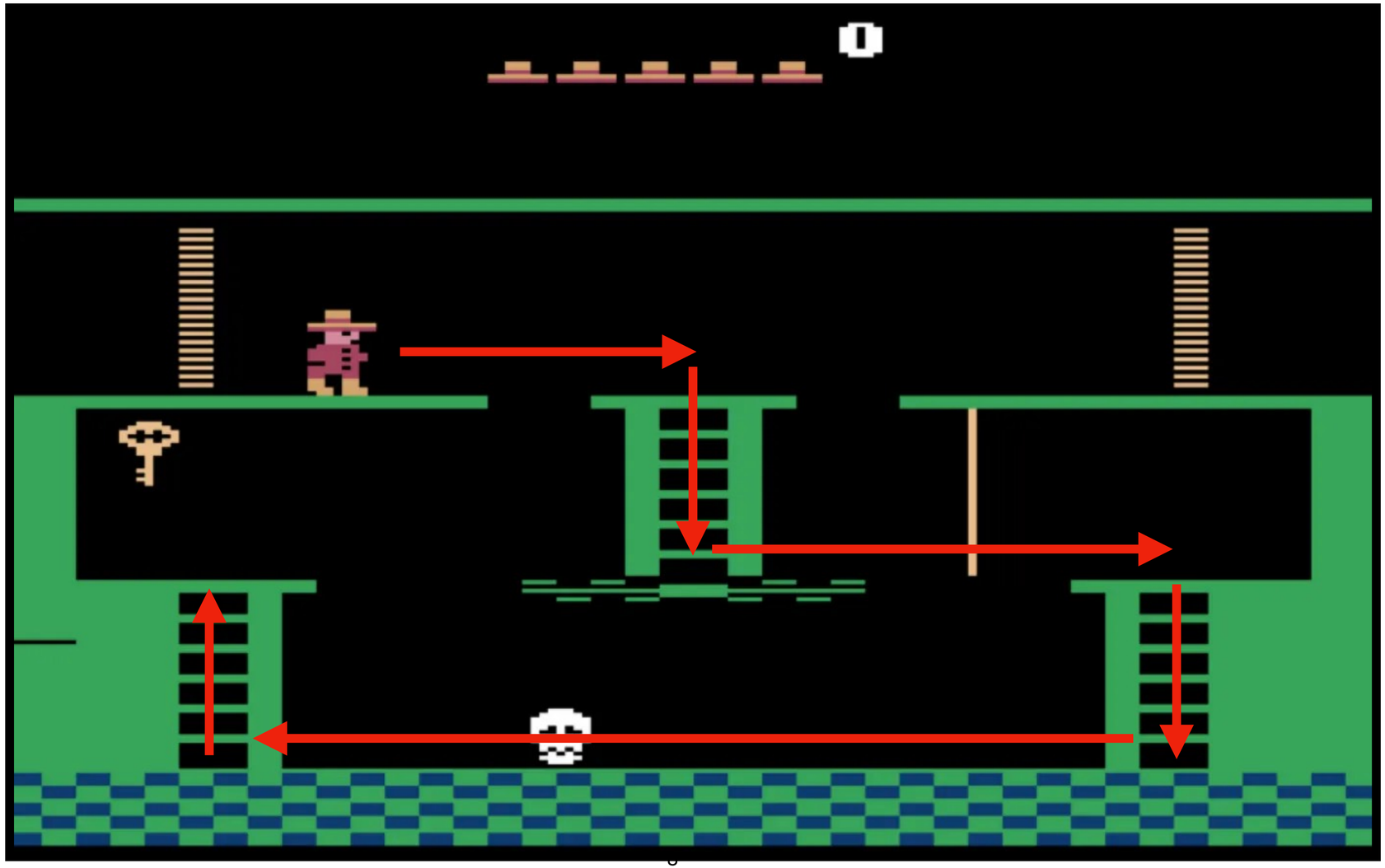
$$J(\pi) = \mathbb{E}_{\pi}\left[r(x, y)\right] - \frac{1}{\eta}\,\mathbb{D}_{\mathrm{KL}}(\pi \,\|\, \pi_{\mathrm{ref}})$$

- Sparse or Delayed Rewards
- Exploration vs. Exploitation
- Credit Assignment
- Sample Inefficiency

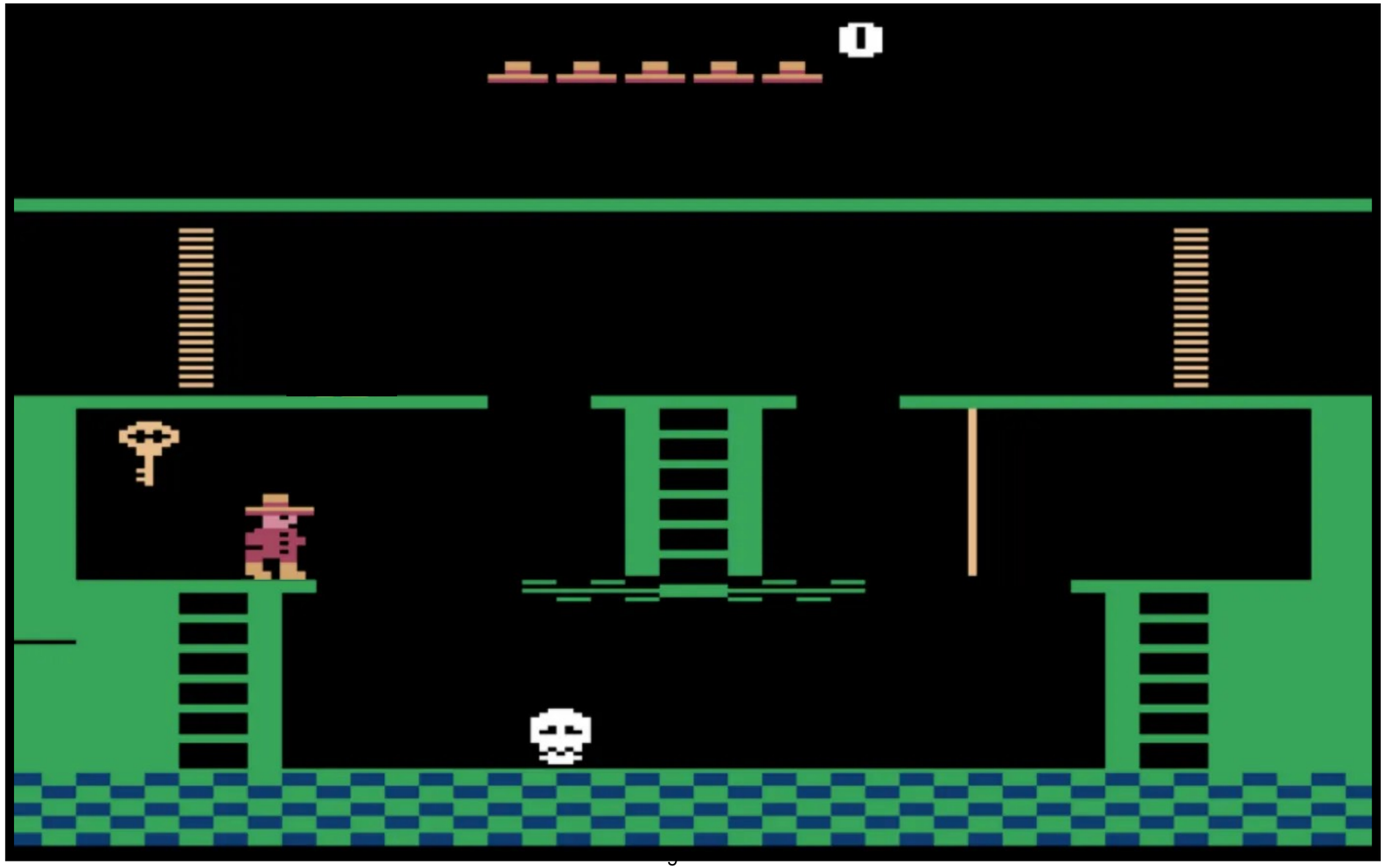
Montezuma Revenge



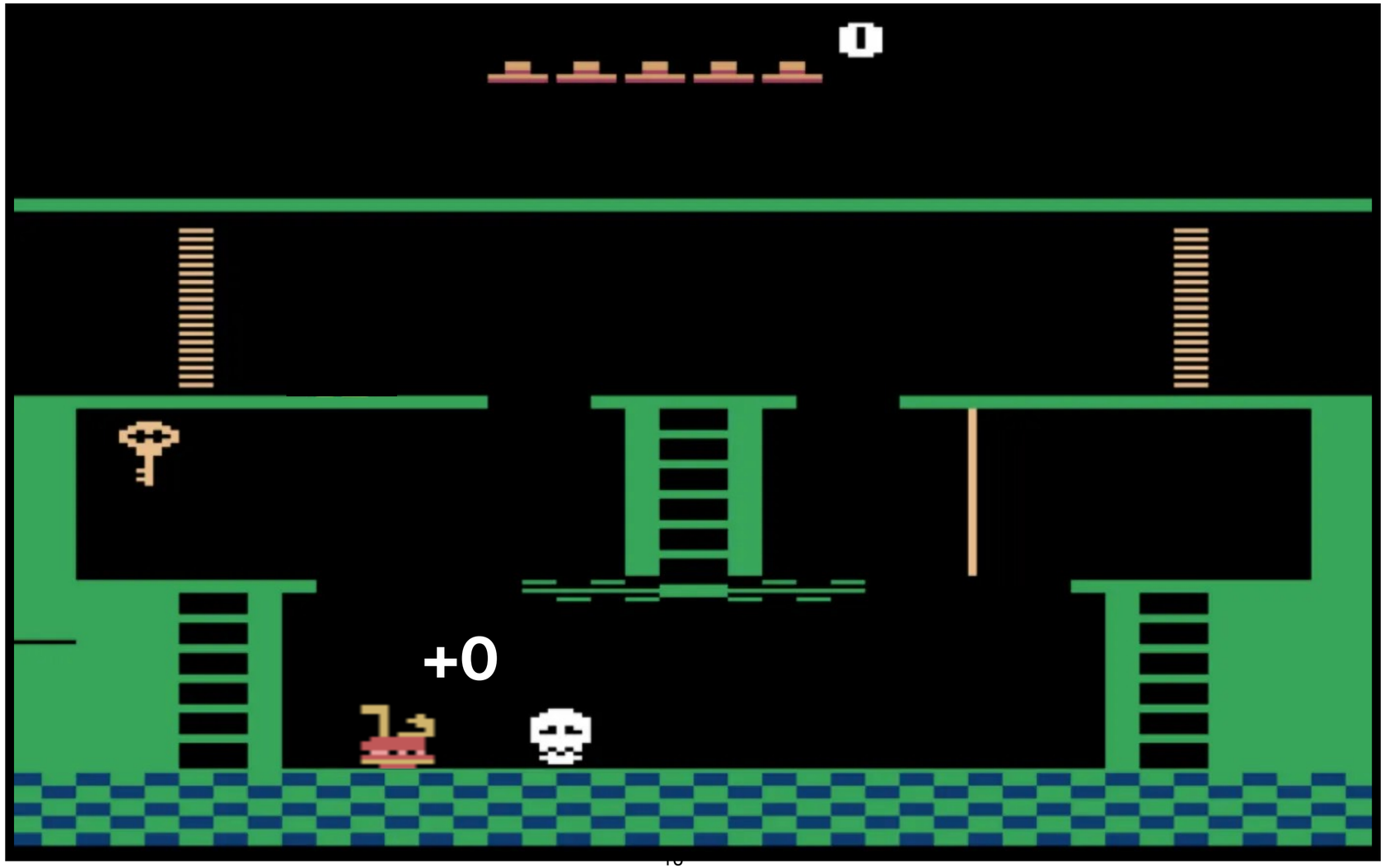
Montezuma Revenge



Montezuma Revenge



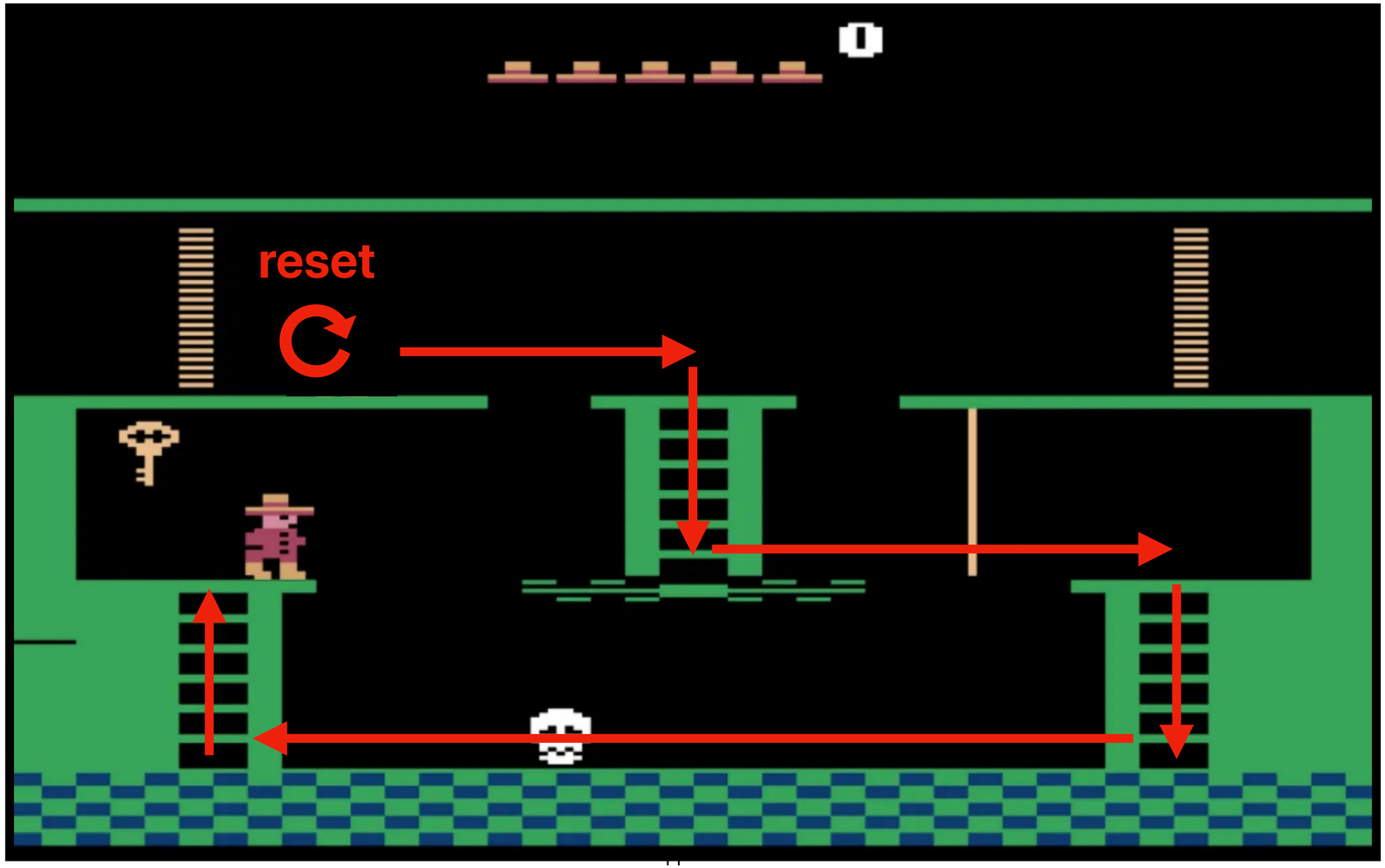
Montezuma Revenge



Montezuma Revenge

Go-Explore: a New Approach for Hard-Exploration Problems by Ecoffet et. al 2019

Learning Montezuma's Revenge from a Single Demonstration by Salimans et. al 2018



How can we develop efficient algorithms for solving the MinxEnt RL objective, assuming access to environment resets?

Outline

- Resetting with reference policy
- Resetting with demonstration data
- Resetting with the current policy

Outline

Resetting with reference policy (to boost exploration)

Reset Property

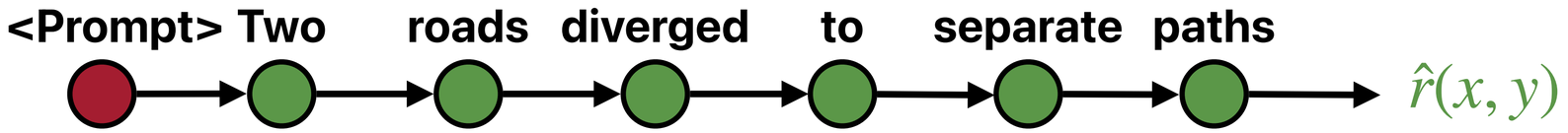
Reset allows us to rollout a policy from partial sentences

?

Inject additional data sources into experience collection

1. Sample a prompt from $x \sim D$

2. Sample a response from $y \sim \pi$



Reset Property

Transition: $P(s' \mid s, a)$ Deterministic

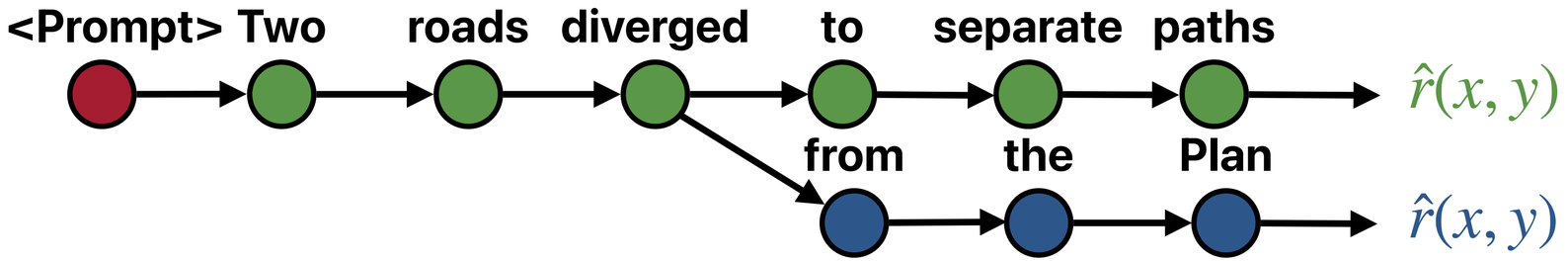
Reset allows us to rollout a policy from partial sentences ?

Inject additional data sources into experience collection

1. Sample a prompt from $x \sim D$

2. Sample a response from $y \sim \pi$

3. Reset and sample a continuation of the response from $y \sim \pi$



Reset Property

Transition: $P(s' \mid s, a)$ Deterministic

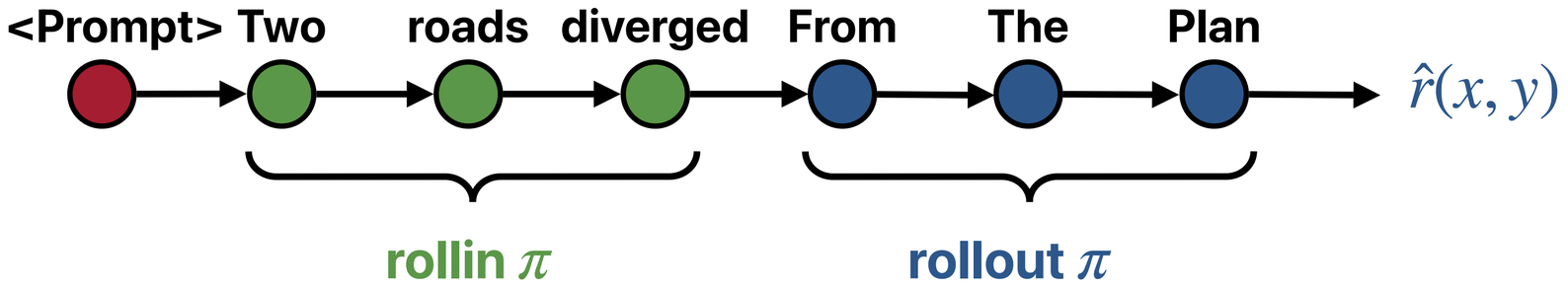
Reset allows us to rollout a policy from partial sentences ?

Inject additional data sources into experience collection

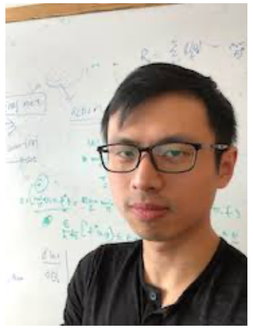
1. Sample a prompt from $x \sim D$

2. Sample a response from $y \sim \pi$

3. Reset and sample a continuation of the response



Learning to Generate Better Than Your Teacher



Wen Sun

Rajkumar Ramamurthy

Dipendra Misra

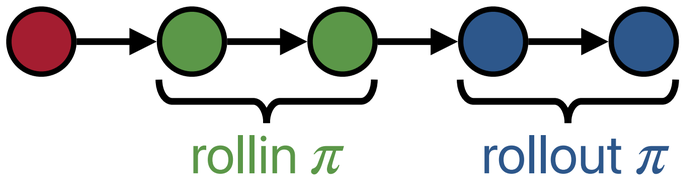
Jonathan D. Chang

Rollin and Rollout
approaches

$$\mathbb{E}_{\pi}\left[\hat{r}(x,y)\right] + \frac{1}{\eta} D_{\mathrm{KL}}(\pi \,||\, \pi_{\mathrm{ref}})$$

Constrains policy to stay near π_{ref}

PPO (RL algorithm)



- Does not leverage problem-specific structure
- Samples prompts $x \sim D$
- Scores action with $\hat{r}(x, y)$

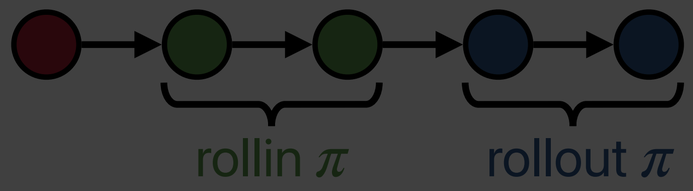
Proximal Policy Optimization, John Schulman et al., 2017

Rollin and Rollout
approaches

$$\mathbb{E}_{\pi}\left[\hat{r}(x,y)\right] + \frac{1}{\eta}\underbrace{D_{\mathrm{KL}}(\pi \,||\, \pi_{\mathrm{ref}})}$$

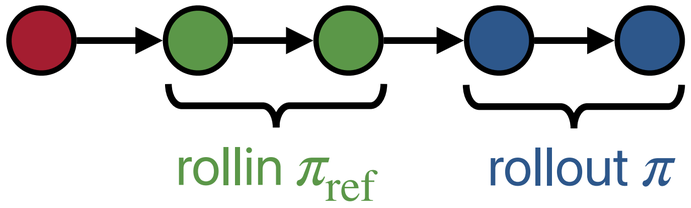
Constrains policy to stay near π_{ref}

PPO (RL algorithm)



- Does not leverage problem-specific structure
- Samples prompts $x \sim D$
- Scores action with $\hat{r}(x, y)$

PPO++



- **Sample prompts from a mixture** $x \sim \beta D + (1-\beta)d^{\pi_{\mathrm{ref}}}$
- Scores actions with $\hat{r}(x, y)$
- **Intuition: Richer initial states boost exploration**

Proximal Policy Optimization, John Schulman et al., 2017

Theory of PPO++

Let $\pi^\star$ be a high quality policy covered by π_{ref}

$$\mathbb{E}_{s\sim\beta\rho^{\pi_{\text{ref}}}+(1-\beta)D}\left[\max_a A^{\pi^t}(s,a)\right] \leq \epsilon$$

$$\overbrace{J(\pi^\star) - J(\pi^t)}^{\text{Performance gap}} \leq O\left(H^2 \max_s \underbrace{\left(\frac{d^{\pi^\star}(s)}{d^{\pi^{\text{ref}}}(s)}\right)}_{\text{Assume bound density ratio and } \pi_{\text{ref}} \text{ provides coverage for } \pi^\star} \epsilon\right)$$

Assume that one-step local improvement over π^t is small

Approximately Optimal Approximate Reinforcement Learning Kakade and Langford 2002

Experimental Setup

Task Statement

Given a reddit post, write a TL;DR (short summary).

Example Post

SUBREDDIT: r/dogs
TITLE: [HELP] Not sure how to deal with new people/dogs and my big ole pup
POST: I have a three year old Dober/Pit mix named Romulus ("Rome" for short). I live with 3 other dogs: a 10 year old labrador, a 2 year old French Bulldog and a 8 year old maltese mix. The four of them get along just fine, Rome and the Frenchie are best best best best friends. He isn't the best at meeting new people, but not ALWAYS....Then, the crux of the matter: I want to have a 4th of July party. Several people want to bring their dogs. I doubt I can say "no dogs allowed" and I don't want to let everyone else bring their dog and make mine stay at day care all day.

Example Human Label

TL;DR: HOW do I introduce new people? HOW do I introduce new dogs? WHAT do I do about 4th of July??

Learning to summarization from human feedback Stiennon et al. 2017

Experimental Setup

Task Statement

Given a reddit post, write a TL;DR (short summary).

Dataset Composition

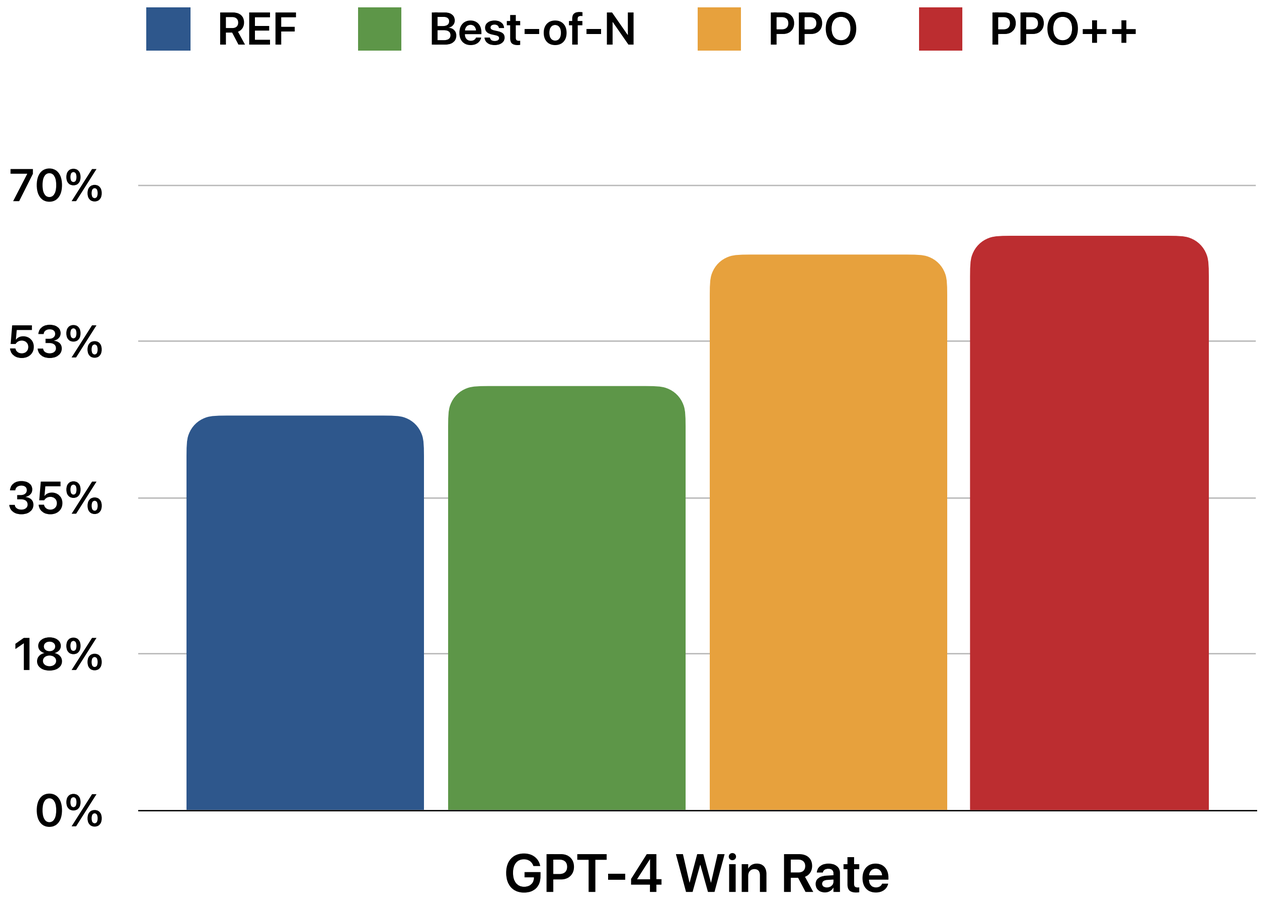
- 210K Prompts total
- 117K Prompts with *Human Labels* } **Used to do RL fine-tuning**
- 93K Prompts with *Human Preference Labels* } **Used to pre-train a reward model**

Experimental Results: TL;DR

GPT4 Winrate Prompt Template

Which of the following summaries does a better job of summarizing the most important points in the given forum Post? FIRST provide a one-sentence comparison of the two summaries, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

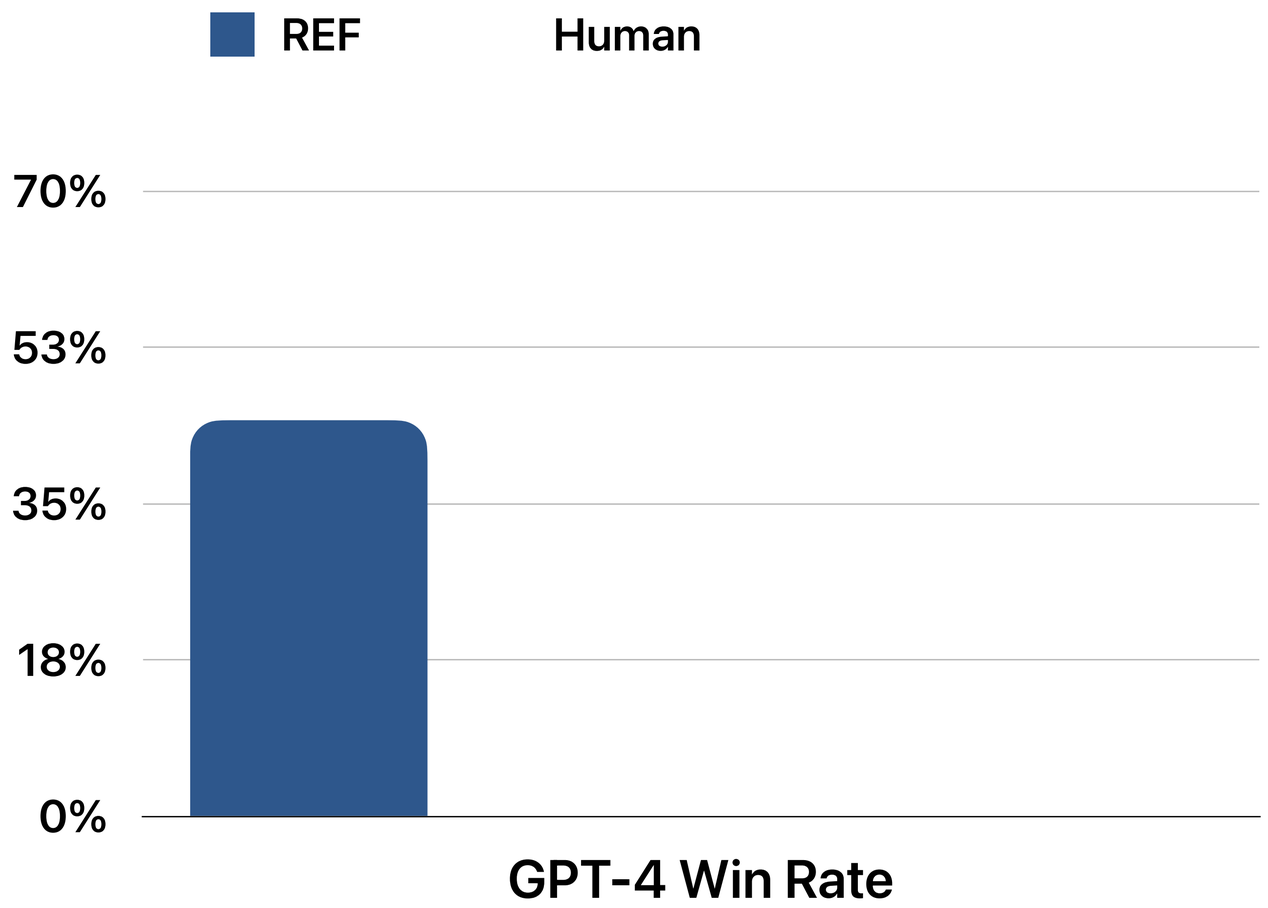
Post: <Post>
A: <TLDR A>
B: <TLDR B>



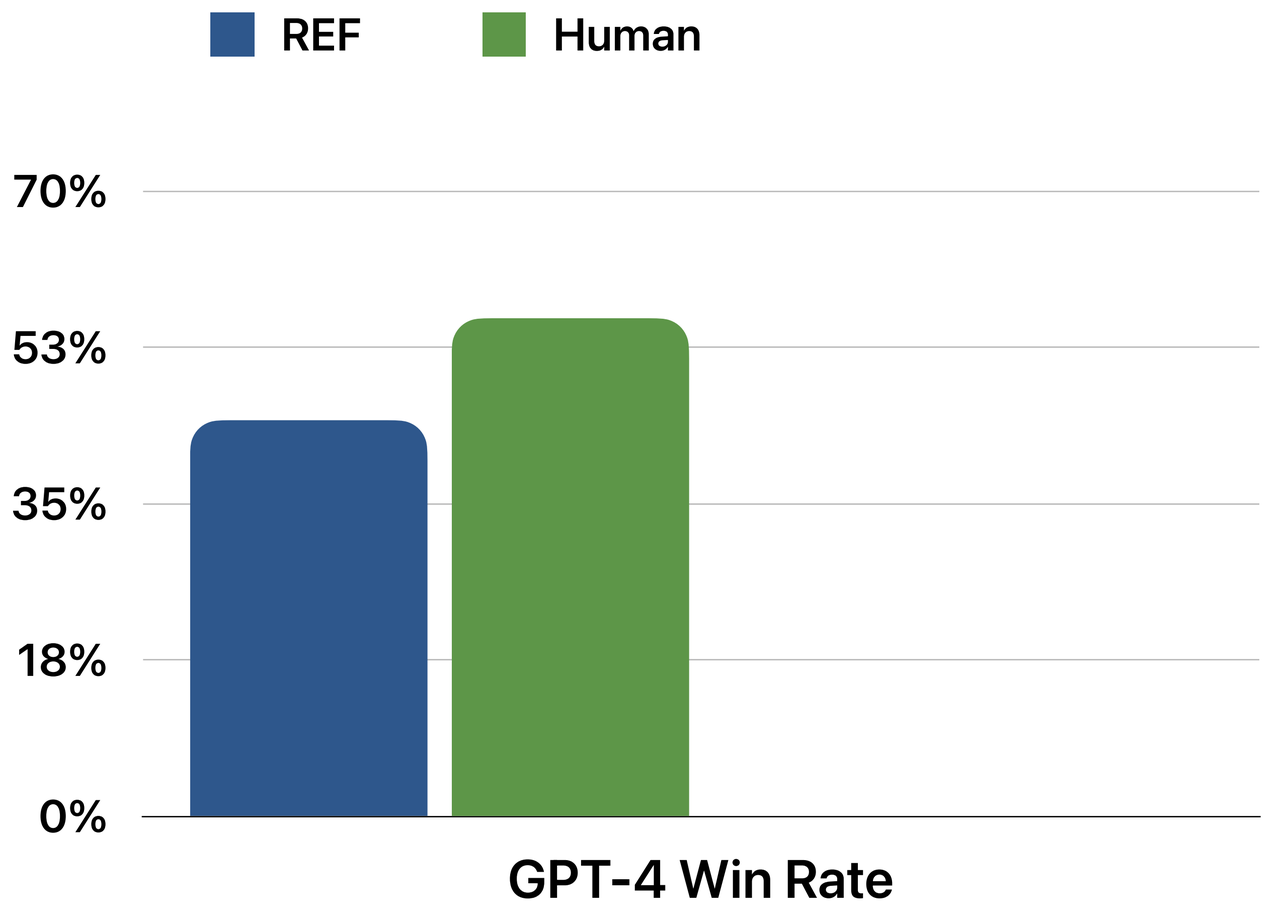
Summary

- PPO^{++} does not leverage a problem-specific structure.
- **The ability to reset is a unique property of MDPs for LLMs.**
- PPO is a simple algorithm that leverages resets.

Can we improve performance by mixing in a different distribution?



Can we improve performance by mixing in a different distribution?



Outline

- **Reset with reference policy**
- **Reset with the demonstration data**

Outline

Reset with the demonstration data (to boost exploration)

Can we improve performance by mixing in a different distribution?

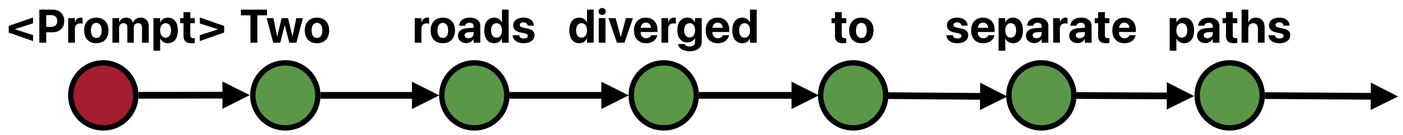
Reset Property

Most text generation tasks have offline label response

!

1. **Sample a prompt from and response from** $(x, y) \sim D$

2. **Reset using the response from** D



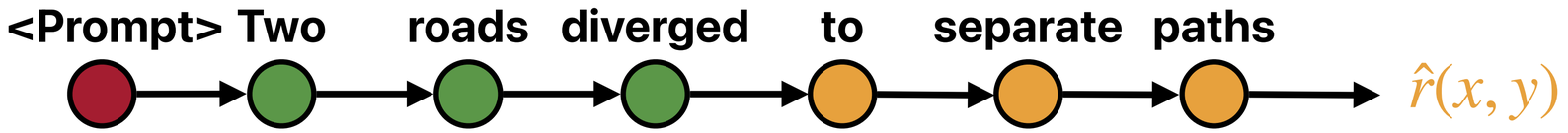
Reset Property

Most text generation tasks have offline label response

1. **Sample a prompt from and response from** $(x, y) \sim D$

2. **Reset using the response from** D

3. **Sample a continuation of the response from** $y \sim \pi$



Dataset Reset Policy Optimization

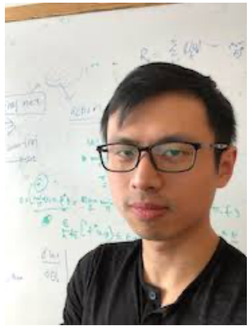
Jonathan D. Chang

Wenhao Zhan

Owen Oertell

Dipendra Misra

Jason Lee

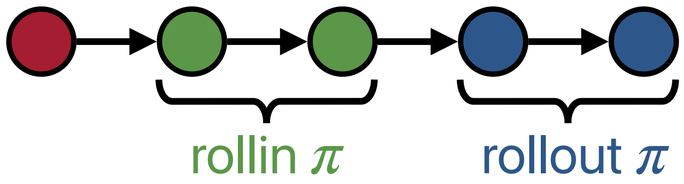


Wen Sun

Rollin and Rollout

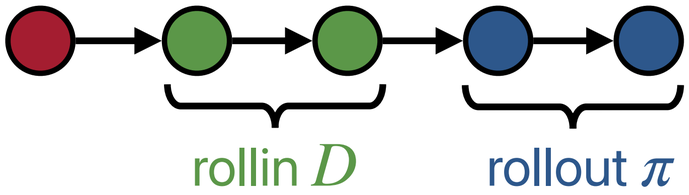
approaches

PPO (RL algorithm)



- Does not leverage problem-specific structure
- Samples prompts $x \sim D$
- Scores action with $\hat{r}(x, y)$

DRPO



- **Sample prompts from a mixture** $x \sim \beta D_x + (1 - \beta) D_{x\|y}$
- Scores actions with $\hat{r}(x, y)$
- **Intuition: Richer initial states boost exploration**

Proximal Policy Optimization, John Schulman et al., 2017

Informal Theory of DR-PO

Informal statement:

When using NPG as the policy optimization oracle, DR-PO learns a policy that is at least as good as any policy covered by the offline data *D*

Coverage assumptions:

$$\frac{d^{\pi^*}(\tau)}{d^{\pi_{\text{ref}}}(\tau)} \leq C_1 < \infty$$

Trajectory-wise density

$$\frac{d^{\pi^*}(x, y)}{d^{\pi_{\text{ref}}}(x, y)} \leq C_2 < \infty$$

State-action sample-wise density

Experimental Setup

Task Statement

Given a reddit post, write a TL;DR (short summary).

Example Post

SUBREDDIT: r/dogs
TITLE: [HELP] Not sure how to deal with new people/dogs and my big ole pup
POST: I have a three year old Dober/Pit mix named Romulus ("Rome" for short). I live with 3 other dogs: a 10 year old labrador, a 2 year old French Bulldog and a 8 year old maltese mix. The four of them get along just fine, Rome and the Frenchie are best best best best friends. He isn't the best at meeting new people, but not ALWAYS....Then, the crux of the matter: I want to have a 4th of July party. Several people want to bring their dogs. I doubt I can say "no dogs allowed" and I don't want to let everyone else bring their dog and make mine stay at day care all day.

Example Human Label

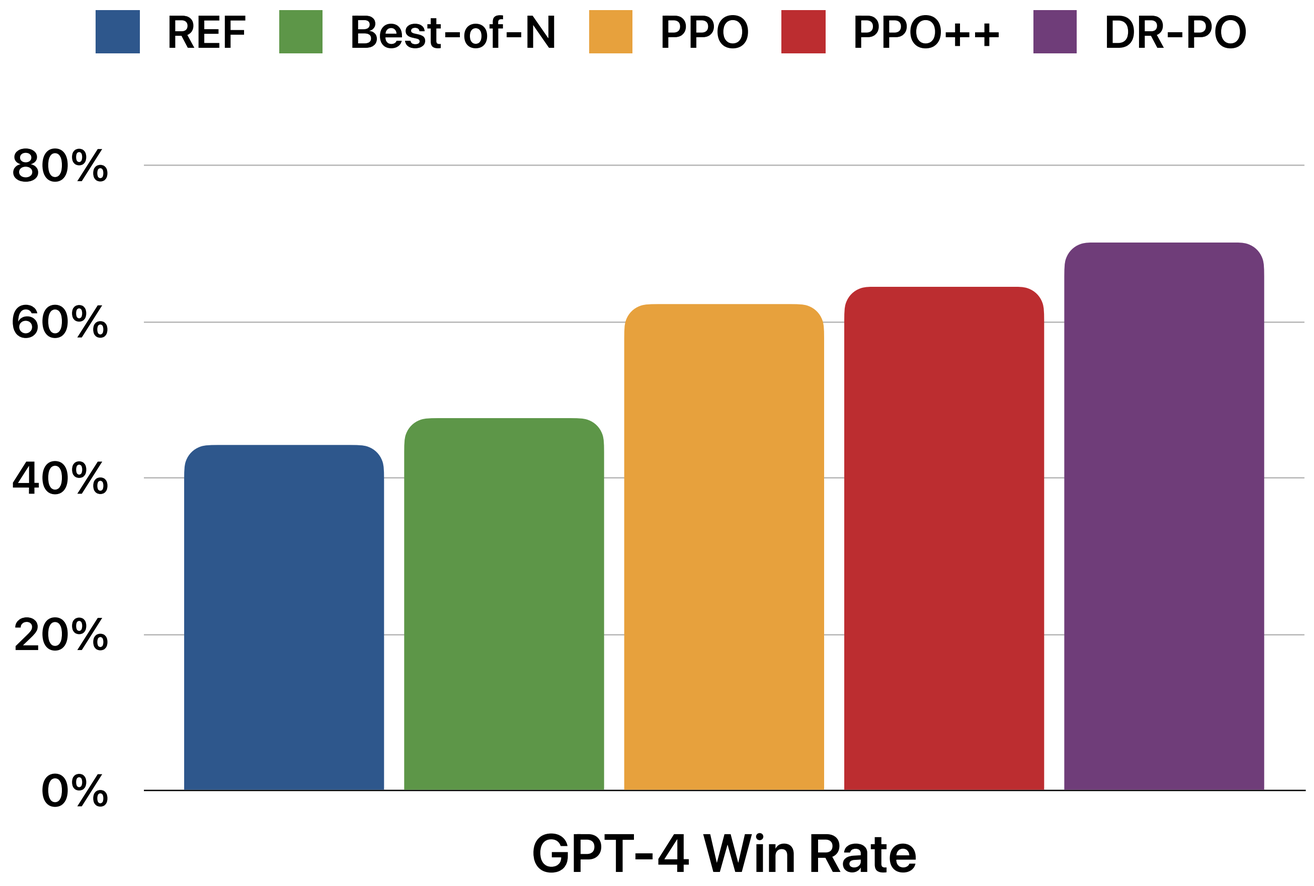
TL;DR: HOW do I introduce new people? HOW do I introduce new dogs? WHAT do I do about 4th of July??

Learning to summarization from human feedback Stiennon et al. 2017

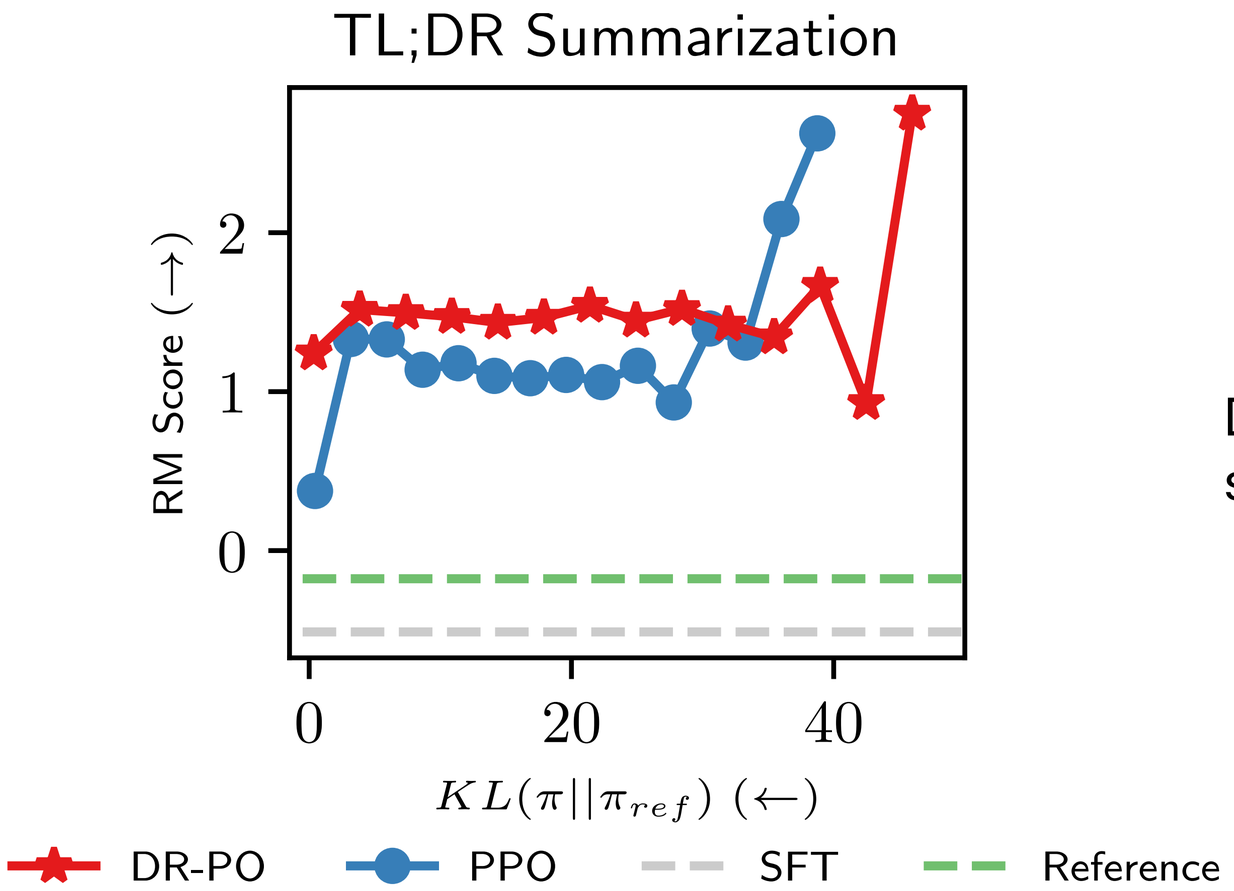
Experimental Results: TL;DR

Takeaways:

1. Algorithms that use resets perform better than those that do not.
2. DR-PO consistently outperforms all baseline algorithms.



Experimental Results: TL;DR



Takeaways:

DR-PO consistently achieves better RM scores with lower KL than PPO.

Experimental Setup

Task Statement

Anthropic's Helpful Harmful task where our model tries to produce an engaging and helpful response to dialogue sequences.

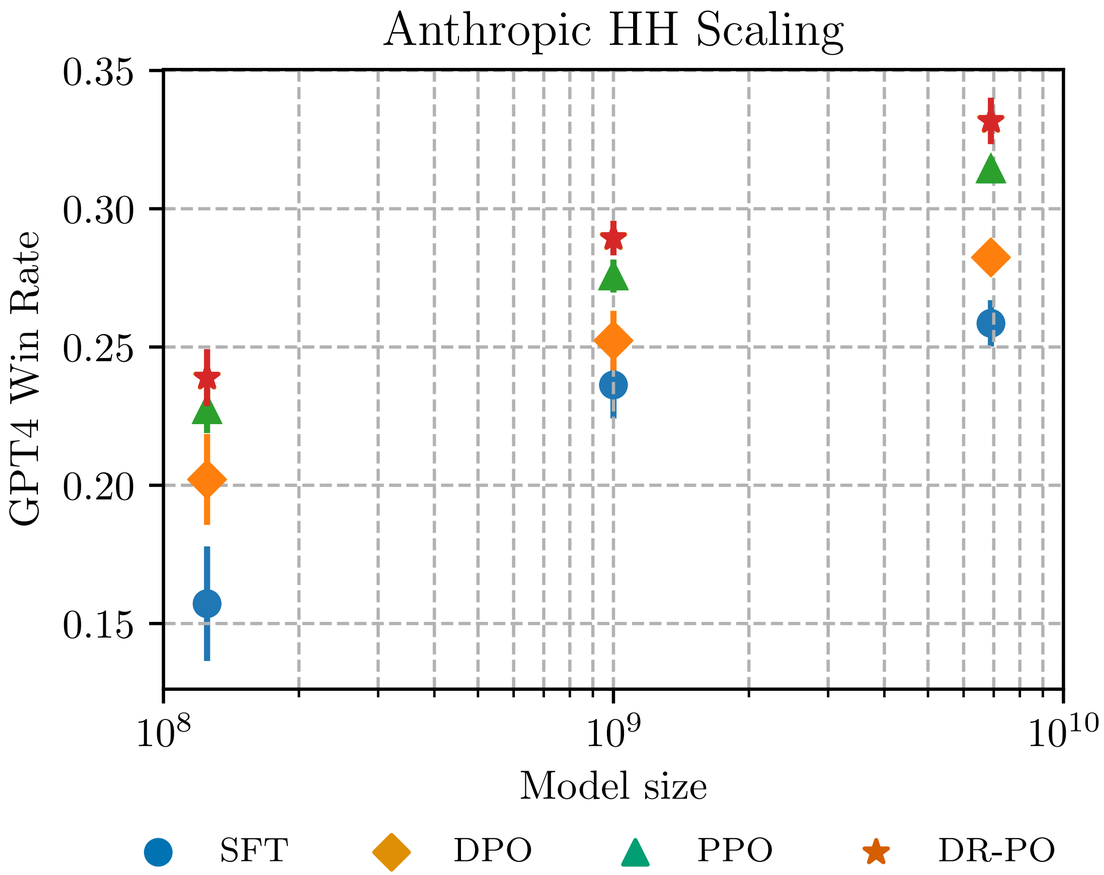
Example Dialogue

Human: What do I do if I crack a molar?

Chosen Assistant: If you cracked a molar, I imagine you're quite concerned, but there's no need to panic, you just need to schedule an appointment with your dentist.

Rejected Assistant: I'm sorry to hear that.

Experimental Results



Takeaways:

1. Online methods outperform offline baselines.
2. DR-PO outperforms all baselines at every model scale.

Summary

- Resetting from offline demonstration data enhances performance.
- DR-PO is provably efficient and improves upon PPO^{++} in theory.
- DR-PO is as simple as PPO and requires no additional computation.

Summary

- Resetting DR-PO from offline demonstration data enhances performance.
- DR-PO is provably efficient and improves upon PPO^{++} in theory.
- DR-PO is as simple as PPO and requires no additional computation.

Outline

- Resetting with reference policy
- Resetting with demonstration data
- Resetting with the current policy

Outline

Resetting with the current policy (to reduce computation)

How can resets be used to make RL algorithms more efficient in computation and memory?

MinXent RL

$$\pi_{i+1} = \arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{\substack{x \sim D \\ y \sim \pi}} \left[r(x, y) \right] - \frac{1}{\eta} D_{\mathrm{KL}}(\pi \parallel \pi_t)$$

$$\forall x, y : \pi_{t+1}(y \mid x) = \frac{\pi_t \exp(\eta r(x, y))}{Z(x)}$$

$$\forall x, y : r(x, y) = \frac{1}{\eta} \left(\ln Z(x) + \ln \left(\frac{\pi_{t+1}(y \mid x)}{\pi_t(y \mid x)} \right) \right)$$

Closed-form solution [Ziebart et al., 2008]:

Rewrite the reward in terms of the policy [Rafailov et al., 2023]:

$$L(r, D) = - \mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(r(x, y_w) - r(x, y_l) \right) \right]$$

Assume reward follows Bradley Terry

Direct Preference Optimization: Your Language Model is Secretly a Reward Model by Rafailov et al. 2023

Maximum Entropy Inverse Reinforcement Learning by Ziebart et al. 2010

MinXent RL

$$\pi_{i+1} = \arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{\substack{x \sim D \\ y \sim \pi}} \left[r(x, y) \right] - \frac{1}{\eta} D_{\mathrm{KL}}(\pi \,\|\, \pi_t)$$

$$\forall x, y : \pi_{t+1}(y \mid x) = \frac{\pi_t \exp(\eta r(x, y))}{Z(x)}$$

$$\forall x, y : r(x, y) = \frac{1}{\eta} \left(\ln Z(x) + \ln \left(\frac{\pi_{t+1}(y \mid x)}{\pi_t(y \mid x)} \right) \right)$$

Closed-form solution [Ziebart et al., 2008]:

Rewrite the reward in terms of the policy [Rafailov et al., 2023]:

Assume nothing about the reward structure, but instead assume access to environment resets.

Direct Preference Optimization: Your Language Model is Secretly a Reward Model by Rafailov et al. 2023

Maximum Entropy Inverse Reinforcement Learning by Ziebart et al. 2010

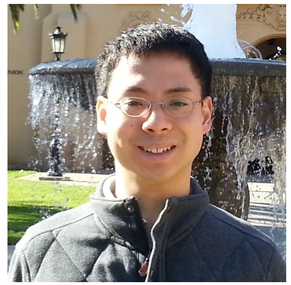
Rebel: Reinforcement learning via regressing relative rewards

Zhaolin Gao

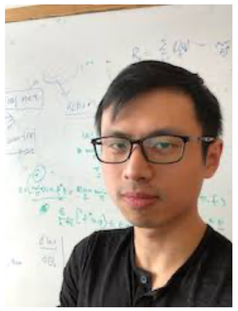
Jonathan D. Chang

Owen Oertell

Jason Lee



Wen Sun



Gokul Swamy

Wenhao Zhan

Thorsten Joachims

J. Andrew Bagnel

REBEL

algorithm overiew

At iteration t with policy π_t

e.g., offline data or reference policy or best-of-N of π_t

1. Sample (hybrid) data using resets:

$$D_t : \{x, y, y'\} \qquad x \sim D, y \sim \pi_t(\cdot \mid x), y' \sim \mu(\cdot \mid x)$$

2. Regressing the relative rewards (least squares regression):

$$\pi_{t+1} = \arg\min_{\pi} \mathbb{E}_{D_t} \left(\frac{1}{\eta} \left(\ln \frac{\pi(y \mid x)}{\pi_t(y \mid x)} - \ln \frac{\pi(y' \mid x)}{\pi_t(y' \mid x)} \right) - \Big(r(x, y) - r(x, y') \Big) \right)^2$$

Predictor

Relative reward

Informal Theory of REBEL

Informal statement:

If we can **solve each regression problem well (in-distribution)**,

$$\pi_{t+1} = \arg\min_{\pi} \mathbb{E}_{D_t}\left(\frac{1}{\eta}\left(\ln\frac{\pi(y \mid x)}{\pi_t(y \mid x)} - \ln\frac{\pi(y' \mid x)}{\pi_t(y' \mid x)}\right) - \Big(r(x,y) - r(x,y')\Big)\right)^2$$

then we can do as well as any policy that is
covered by the training data distributions

$$\forall t, \max_{x,y} \frac{\pi^\star(y \mid x)}{\pi_t(y \mid x) + \mu(y \mid x)} \leq C < \infty$$

Experimental Setup

Task Statement

Given a reddit post, write a TL;DR (short summary).

Example Post

SUBREDDIT: r/dogs
TITLE: [HELP] Not sure how to deal with new people/dogs and my big ole pup
POST: I have a three year old Dober/Pit mix named Romulus ("Rome" for short). I live with 3 other dogs: a 10 year old labrador, a 2 year old French Bulldog and a 8 year old maltese mix. The four of them get along just fine, Rome and the Frenchie are best best best best friends. He isn't the best at meeting new people, but not ALWAYS....Then, the crux of the matter: I want to have a 4th of July party. Several people want to bring their dogs. I doubt I can say "no dogs allowed" and I don't want to let everyone else bring their dog and make mine stay at day care all day.

Example Human Label

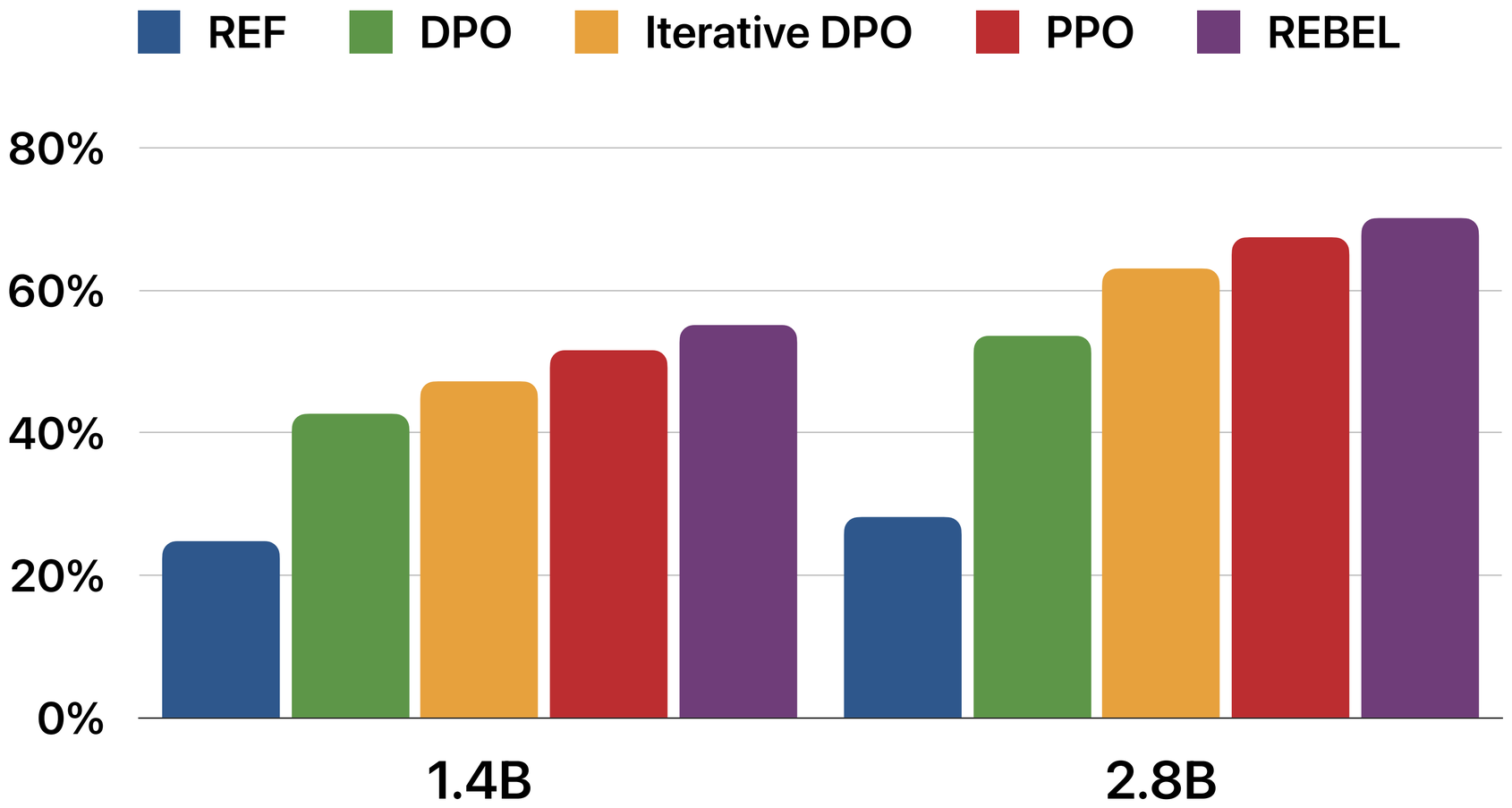
TL;DR: HOW do I introduce new people? HOW do I introduce new dogs? WHAT do I do about 4th of July??

Learning to summarization from human feedback Stiennon et al. 2017

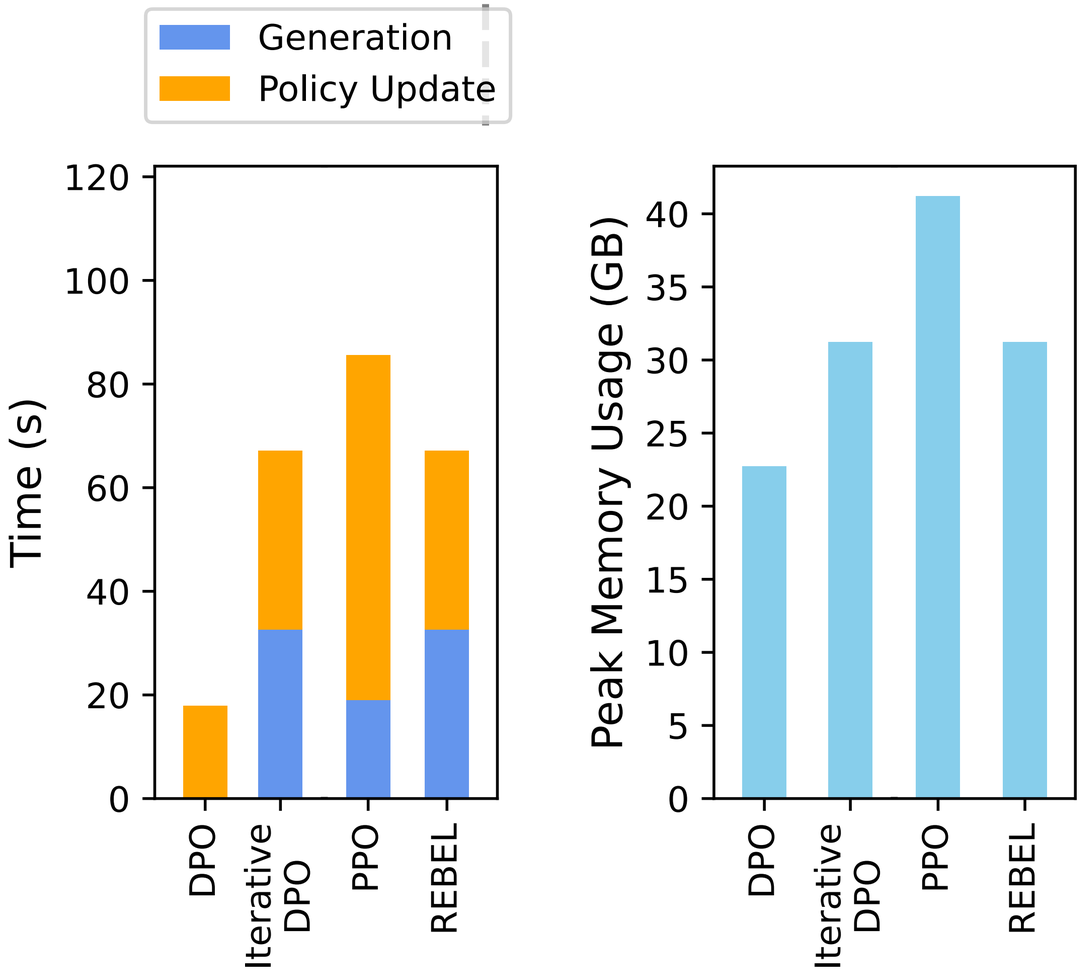
Experimental Results: TL;DR

Takeaways:

1. Online methods outperform offline methods
2. REBEL consistently outperforms all baseline methods across scales.



Experimental Results



Takeaways:

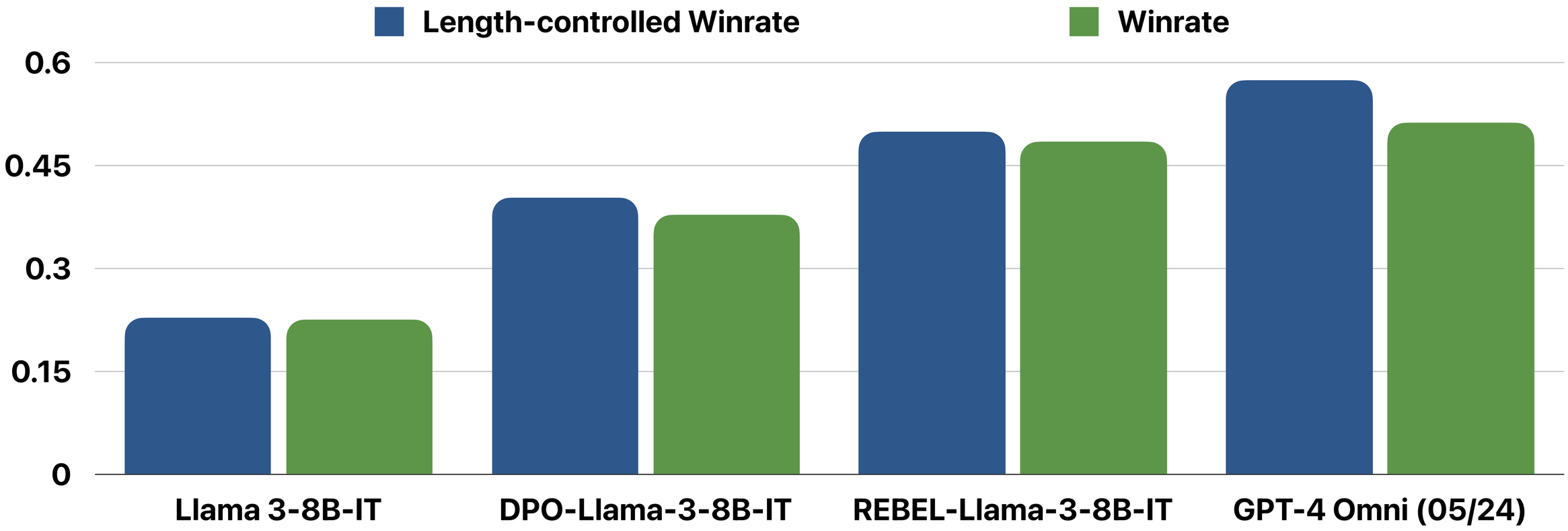
1. **Offline methods** are more efficient but yield lower win rates.
2. REBEL outperforms PPO in both win rate and resource efficiency.
3. REBEL outperforms iterative DPO in win rate while matching its efficiency.

Scaling to larger model (8B) on more modern benchmarks

Experimental Results

Fine-tuning Llama 3-8B model for general chat

Dataset: ultrafeedback [Cui et al]; **Reward Model:** ArMo [Wang et al]

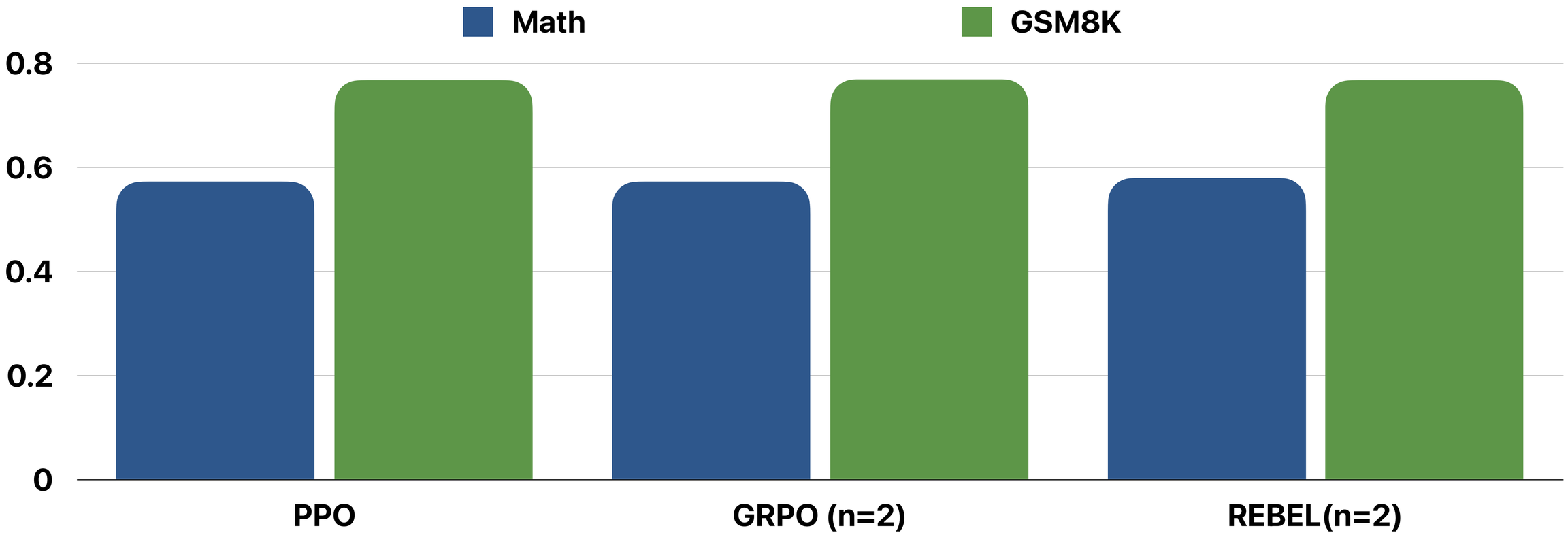


Scaling to Reasoning Tasks

Experimental Results

Fine-tuning Qwen-2.5B model

Dataset: Math and GSM8K; **Reward Model:** Verifiable Reward



Summary

- REBEL reduces the RL problem to a sequence of relative reward regression tasks.
- Empirically, REBEL outperforms PPO in performance, computational efficiency, and memory usage.
- REBEL achieves strong results on standard LLM benchmarks.

How can we develop efficient algorithms for solving the MinxEnt RL objective, assuming access to environment resets?

- Resetting with reference policy (to boost exploration)
- Resetting with demonstration data (to boost exploration)
- Resetting with the current policy (to reduce computation)

Acknowledgement