



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

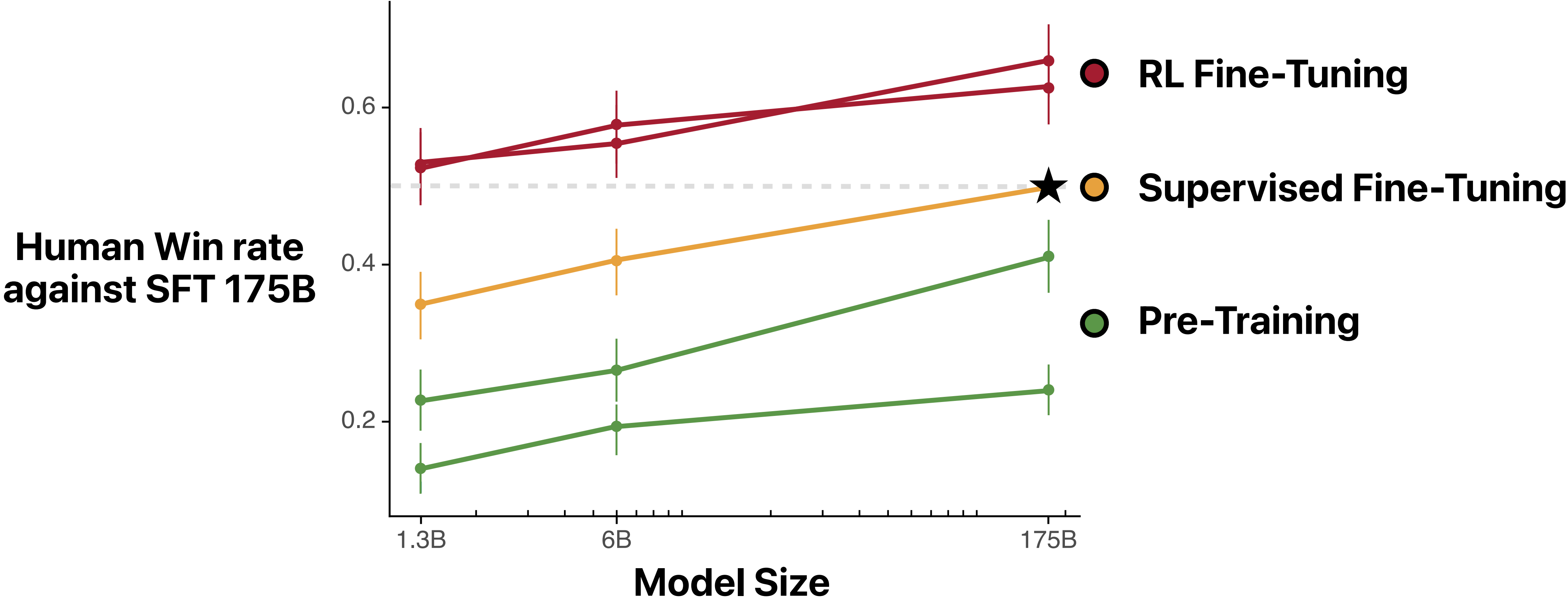
"A large language model is a deep learning model trained on massive text datasets to understand and generate human language."



Large Language Model

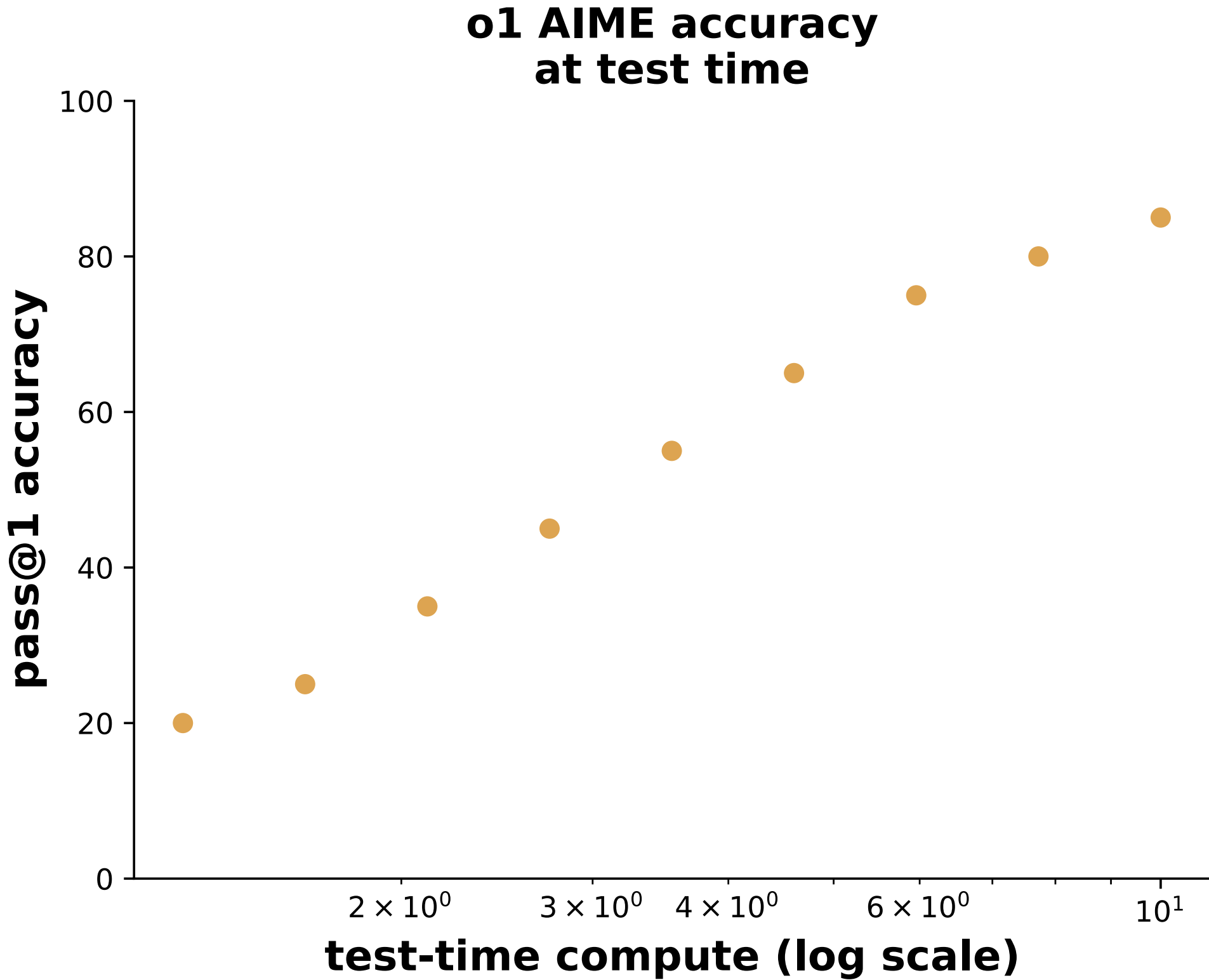
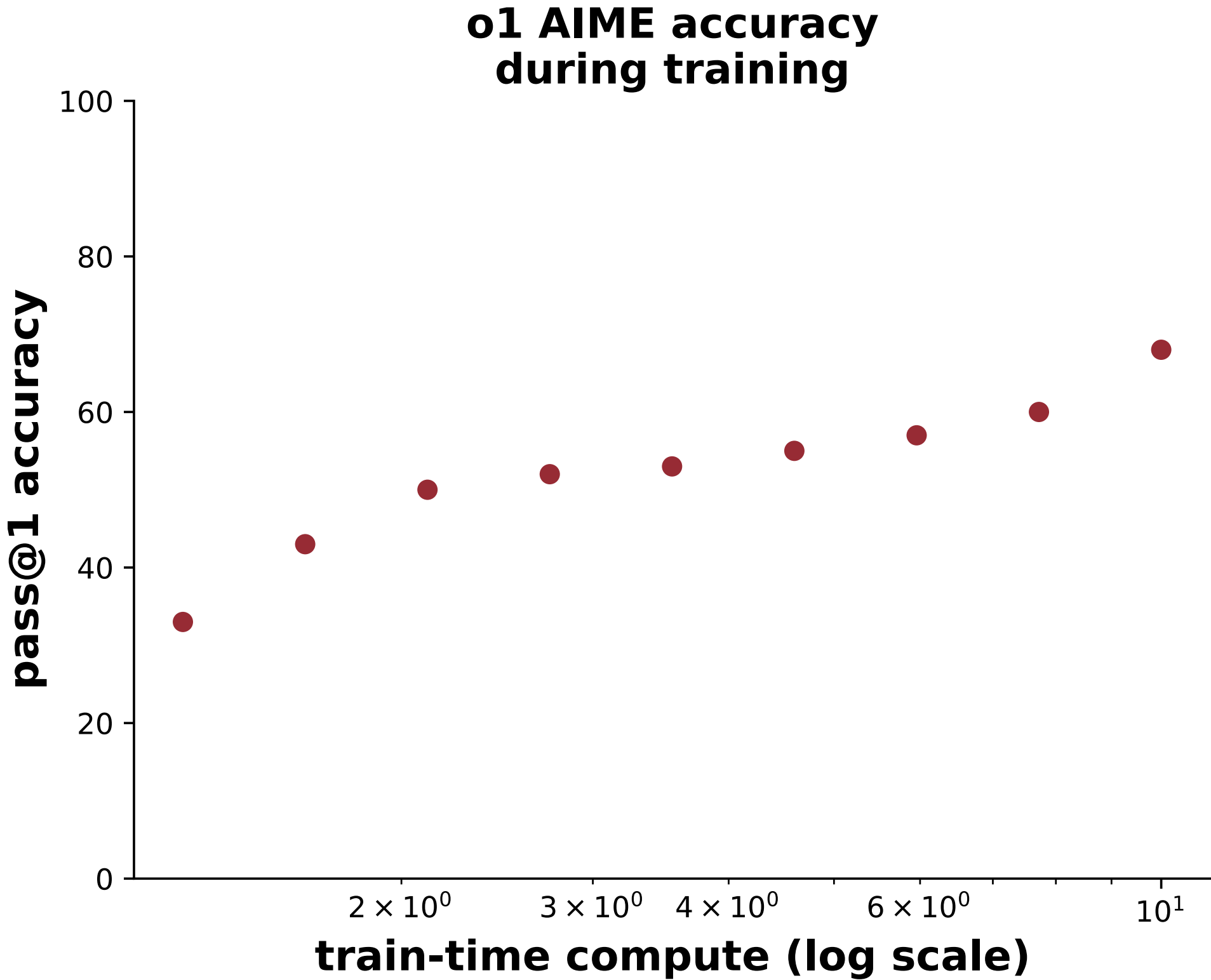
How do *different* objectives shape LLM behavior?

Learning from Human Feedback



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

$$J(\pi) = \underbrace{\mathbb{E}_{\pi} [r(x, y)]}_{\text{Maximize reward}} - \frac{1}{\eta} \underbrace{\mathbb{D}_{\text{KL}}(\pi \parallel \pi_{\text{ref}})}_{\text{Minimal deviation}}$$

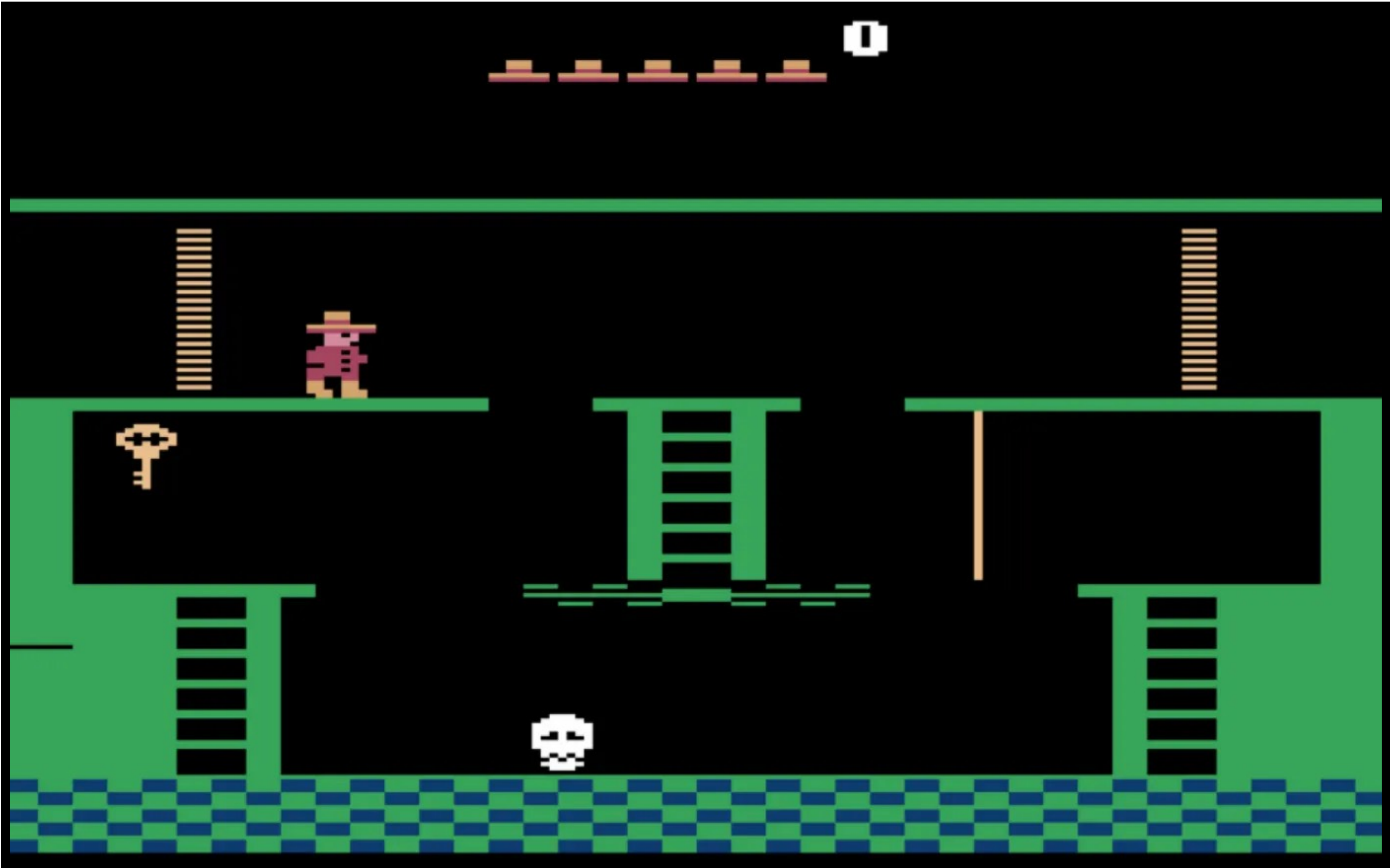
Maximize reward under Minimal deviation

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

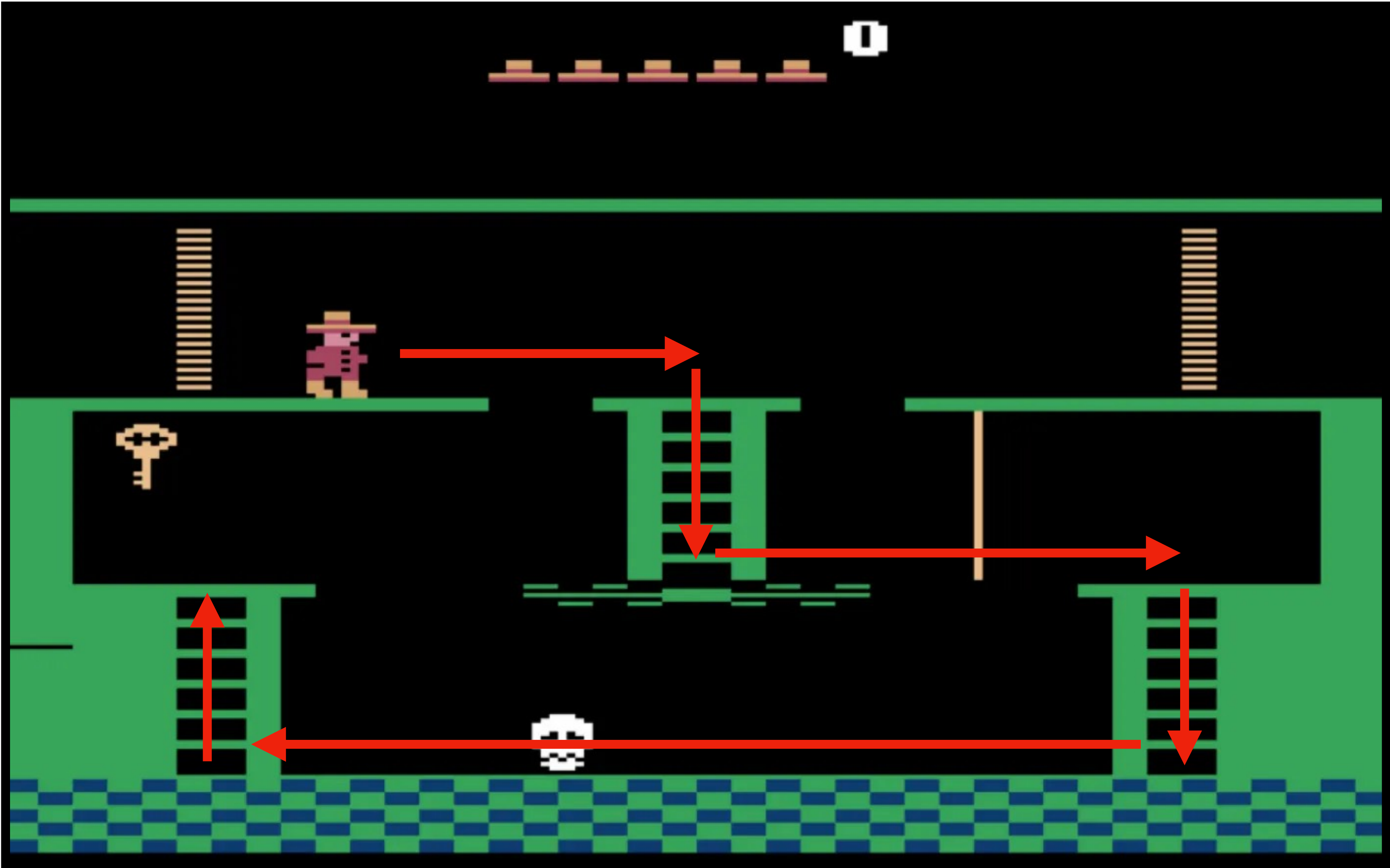

$$J(\pi) = \mathbb{E}_{\pi} [r(x, y)] - \frac{1}{\eta} \mathbb{D}_{\text{KL}}(\pi \parallel \pi_{\text{ref}})$$

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

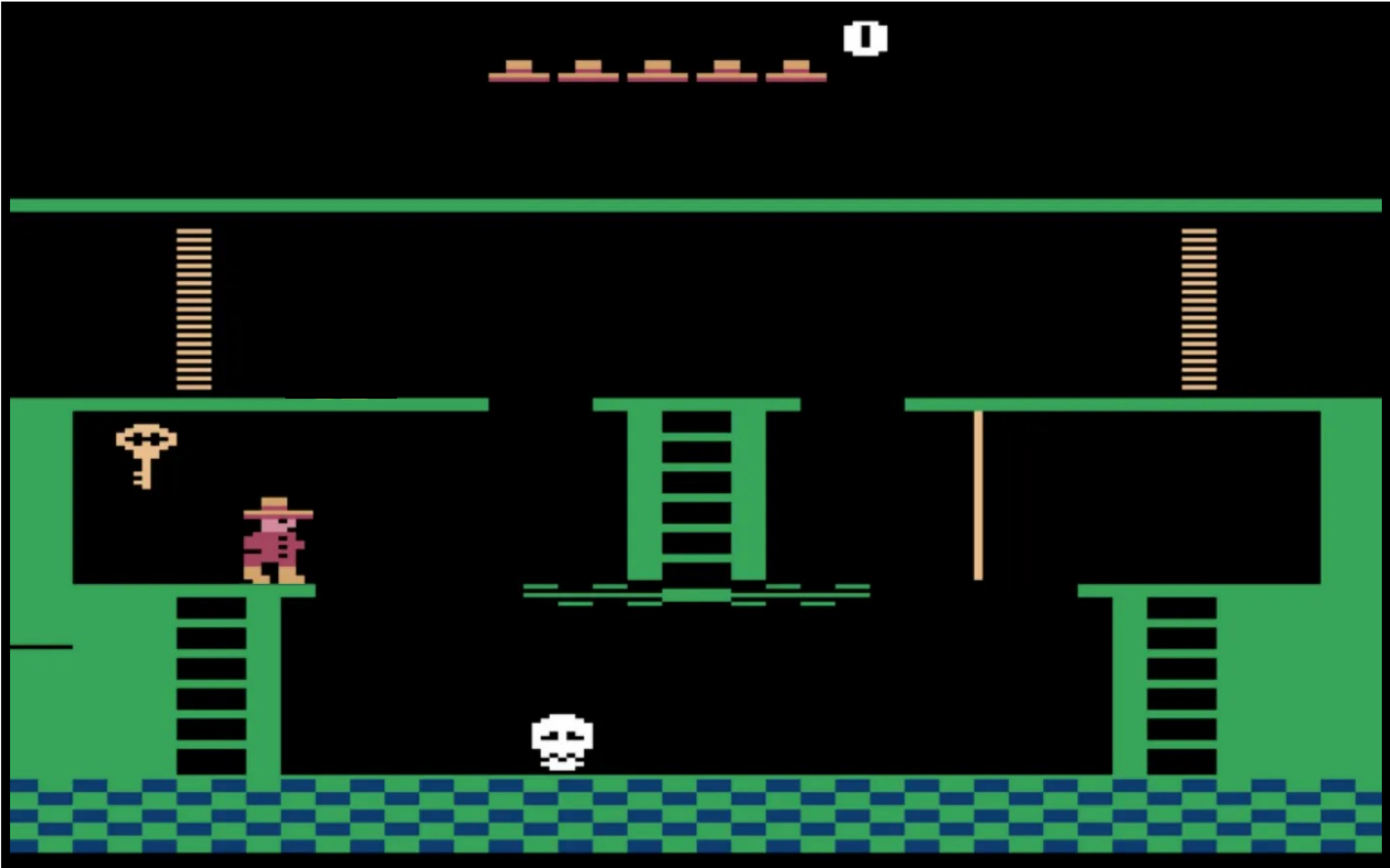
Montezuma Revenge



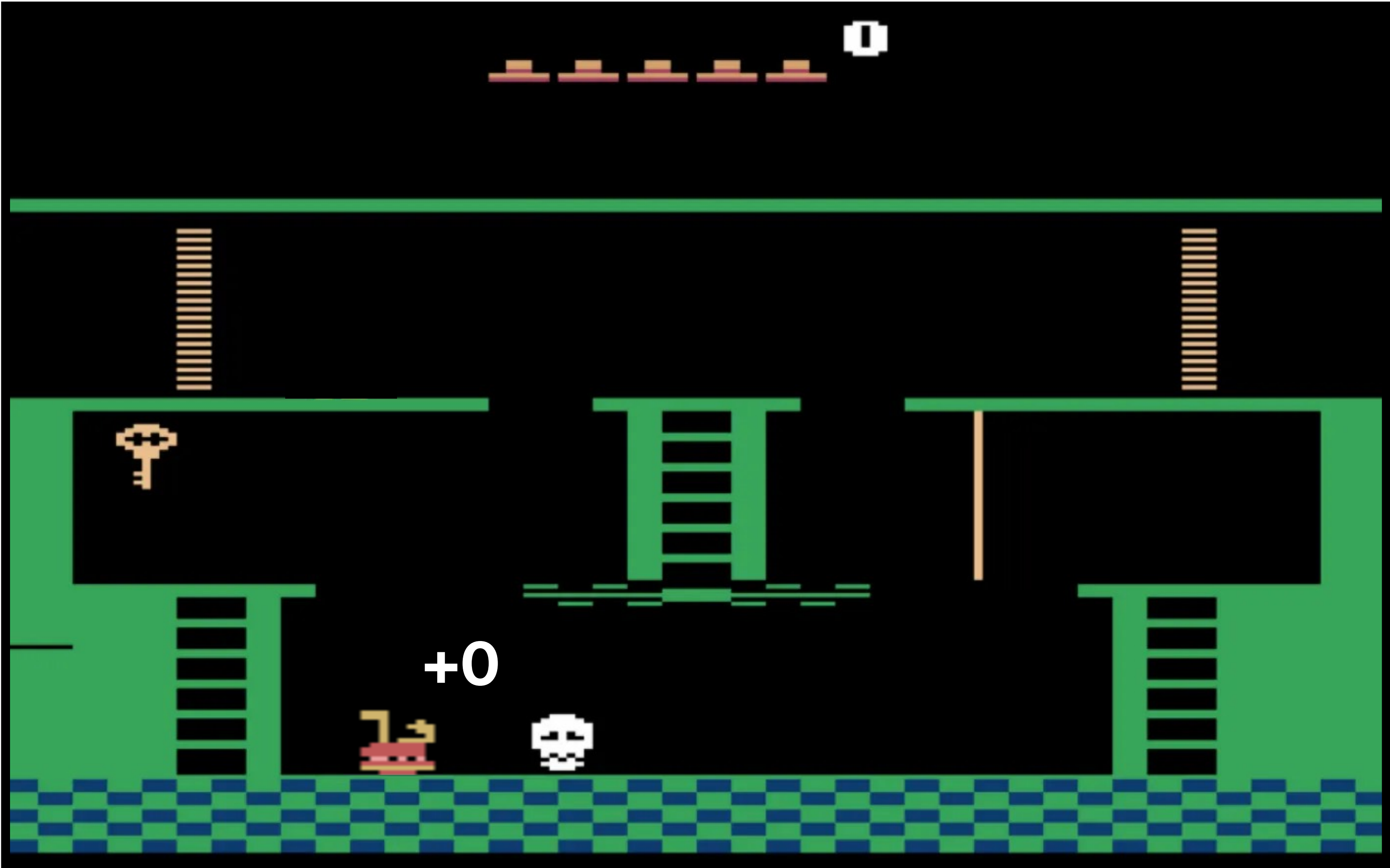
Montezuma Revenge



Montezuma Revenge



Montezuma Revenge



Montezuma Revenge

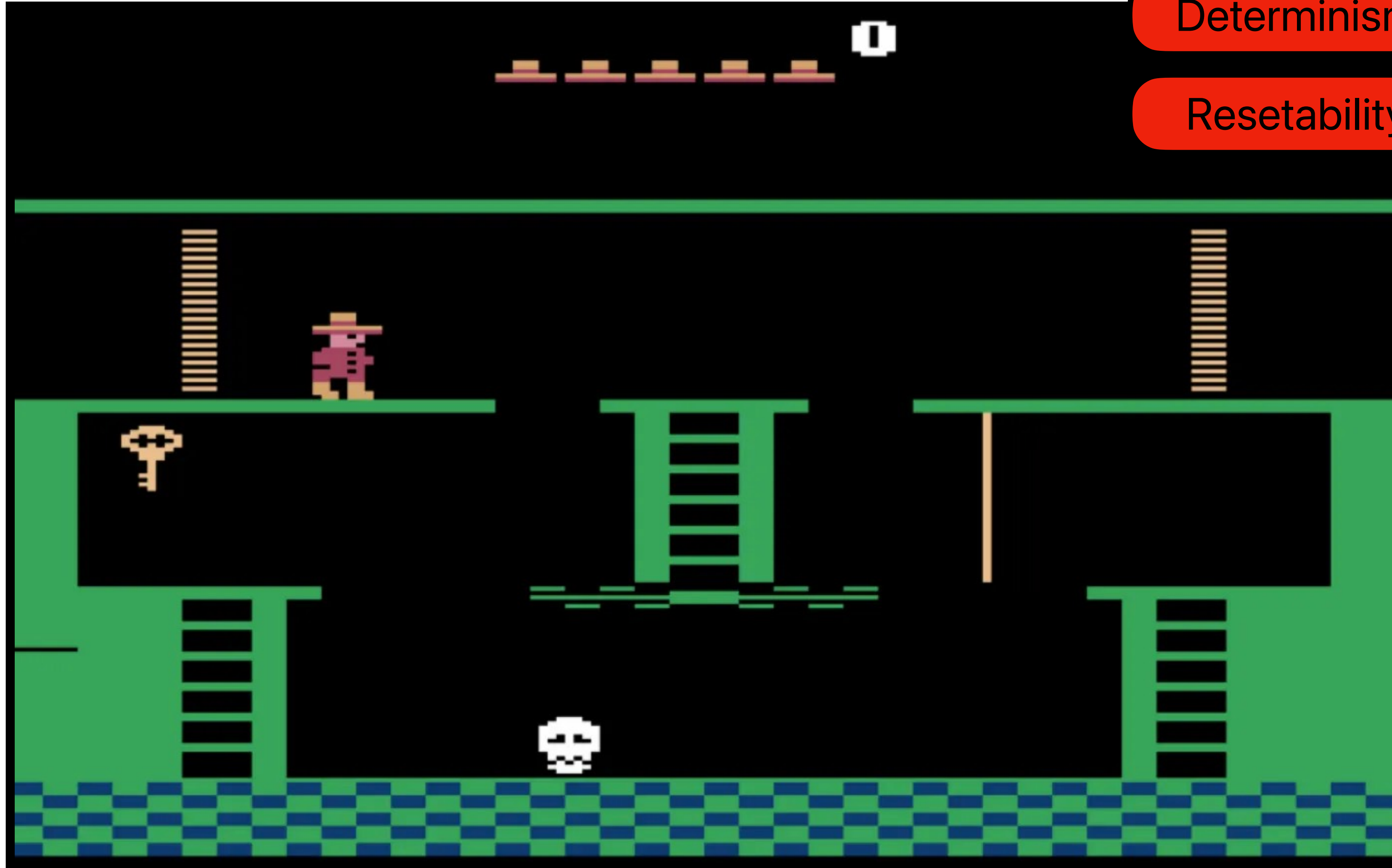
Go-Explore: a New Approach for Hard-Exploration Problems by Ecoffet et. al 2019

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

Key Environment Properties:

Determinism

Resetability



Montezuma Revenge

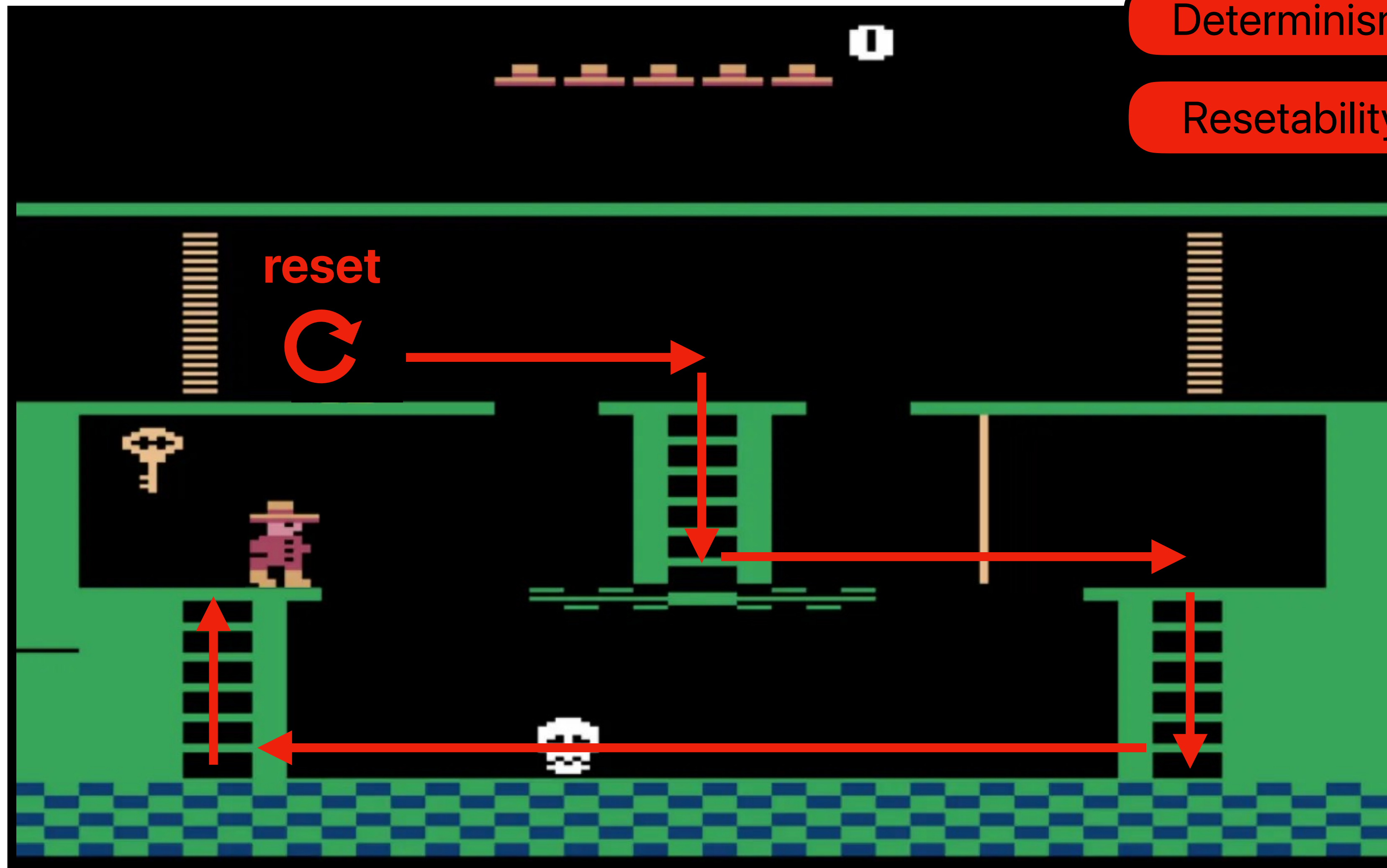
Go-Explore: a New Approach for Hard-Exploration Problems by Ecoffet et. al 2019

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

Key Environment Properties:

Determinism

Resetability



How can environment resets be utilized to solve the MinxEnt RL objective efficiently?

Outline

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

Outline

Resetting with reference policy (to boost exploration)

Reset Property

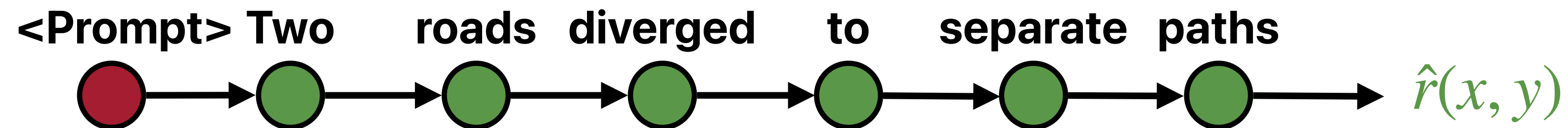
?

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: $P(s' | s, a)$

Deterministic

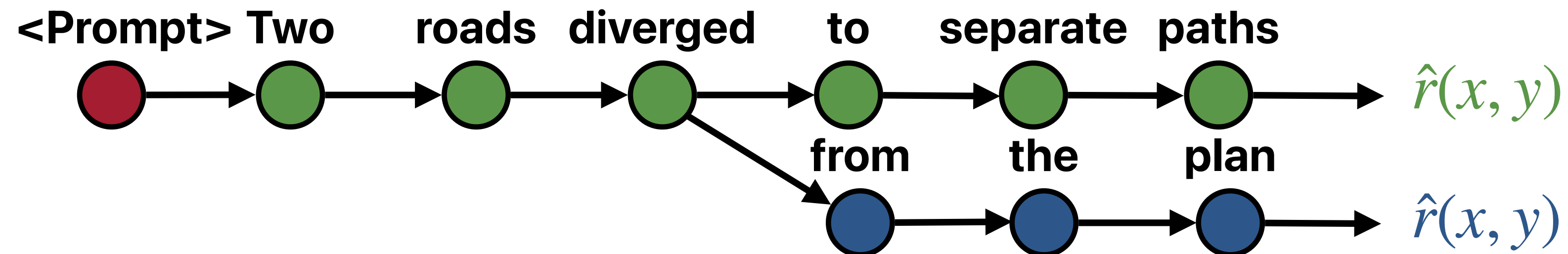
Reset allows us to rollout a policy from partial sentences

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: $P(s' | s, a)$

Deterministic

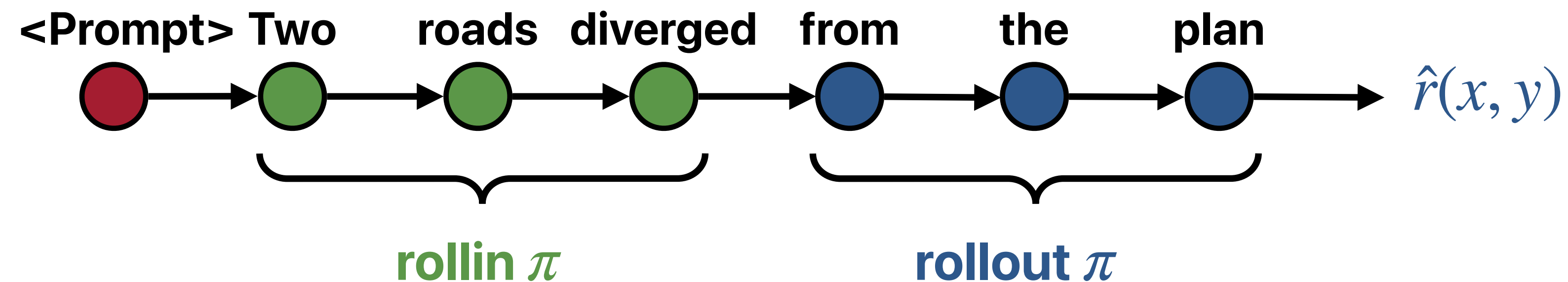
Reset allows us to rollout a policy from partial sentences

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 $y \sim \pi$



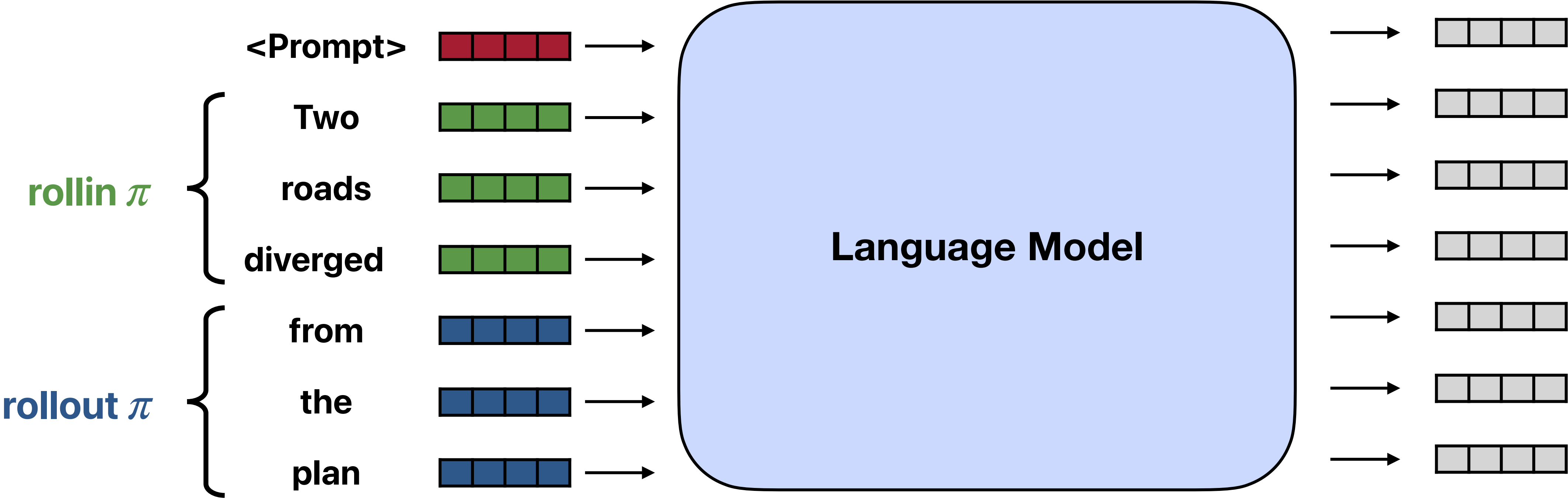
Reset Property

Transition: $P(s' | s, a)$

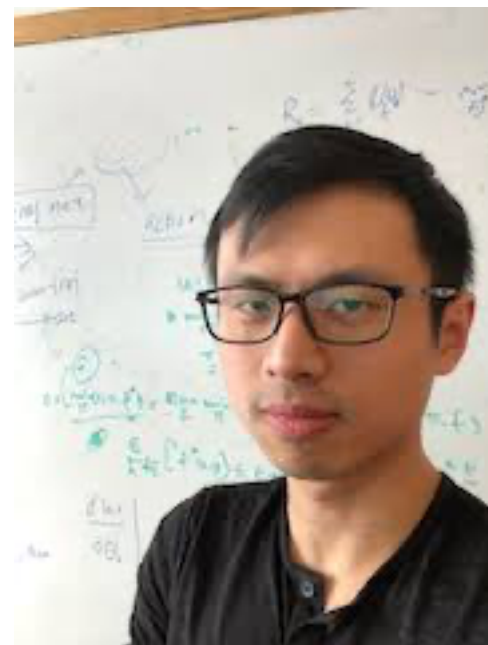
Deterministic

Reset allows us to rollout a policy from partial sentences

Inject additional data sources into experience collection



Learning to Generate Better Than Your Teacher



Wen Sun



Rajkumar Ramamurthy



Dipendra Misra



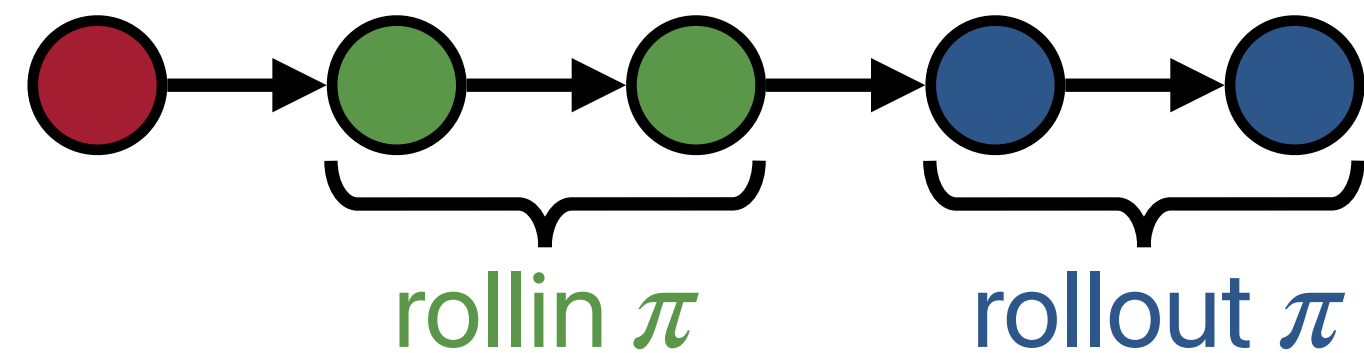
Jonathan D. Chang

Rollin and Rollout approaches

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

Constrains policy to stay near π_{ref}

PPO (RL algorithm)



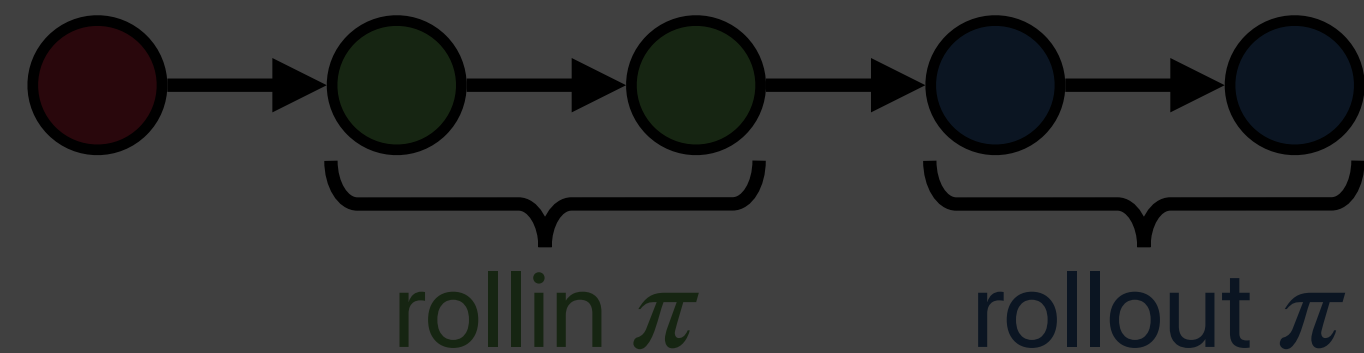
- Does not leverage problem-specific structure
- Samples prompts $x \sim D$
- Scores action with $\hat{r}(x, y)$

Rollin and Rollout approaches

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

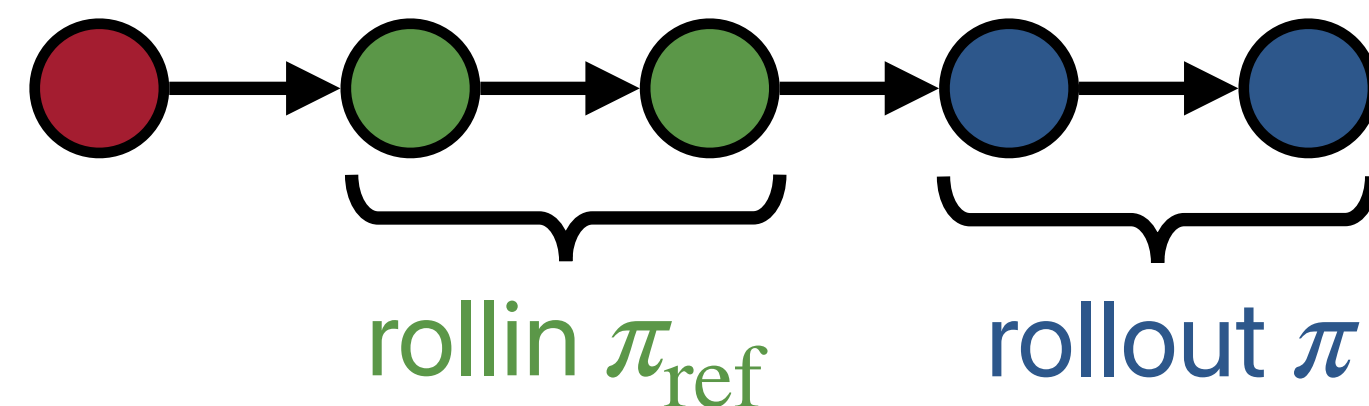
Constrains policy to stay near π_{ref}

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

Theory of PPO++

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

$$\underbrace{J(\pi^\star) - J(\pi^t)}_{\text{Performance gap}} \leq O \left(H^2 \underbrace{\max_s \left(\frac{d^{\pi^\star}(s)}{d^{\pi_{\text{ref}}}(s)} \right)}_{\substack{\text{Assume bound density ratio and} \\ \pi_{\text{ref}} \text{ provides coverage for } \pi^\star}} \underbrace{\mathbb{E}_{s \sim \beta D + (1-\beta)d^{\pi_{\text{ref}}}} \left[\max_a A^{\pi^t}(s, a) \right]}_{\substack{\text{Assume that one-step local} \\ \text{improvement over } \pi^t \text{ is small}}} \epsilon \right)$$

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??

Experimental Setup

Task Statement

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

Dataset Composition

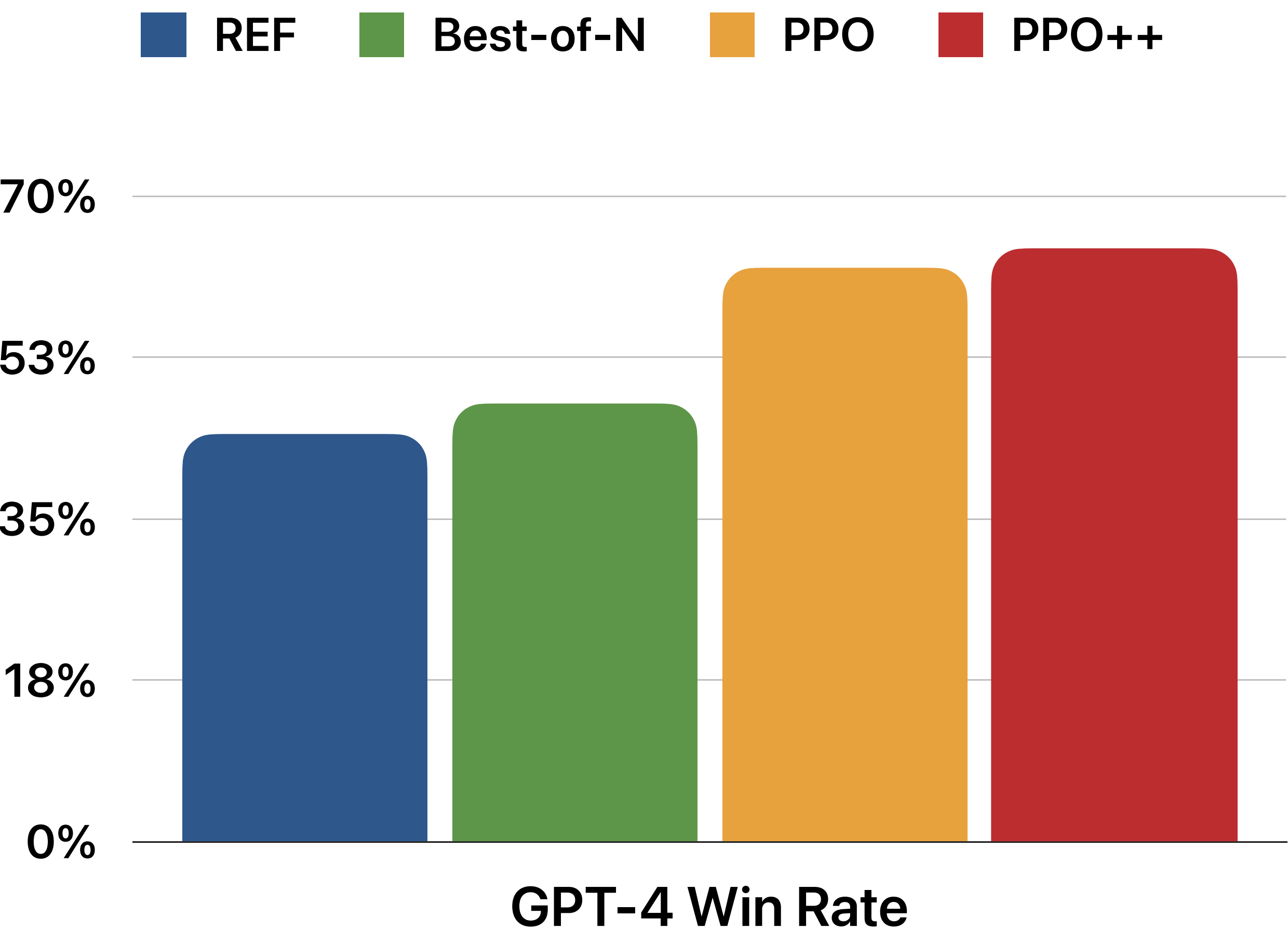
- 210K Prompts total
- 117K Prompts with *Human Labels* } Used to do RL fine-tuning
- 93K Prompts with *Human Preference Labels* } Used to pre-train a reward model

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.

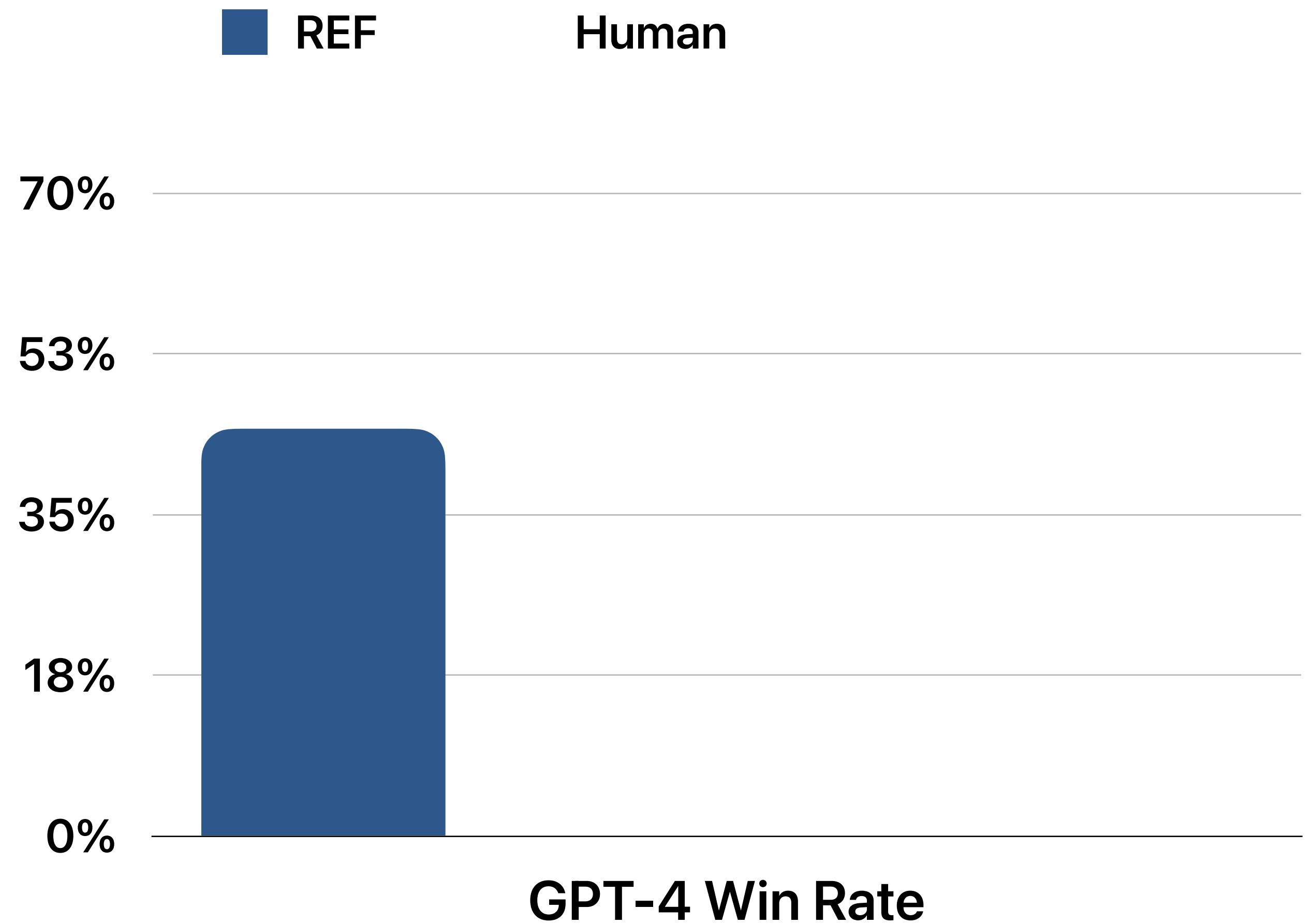
Post: <Post>
A: <TLDR A>
B: <TLDR B>



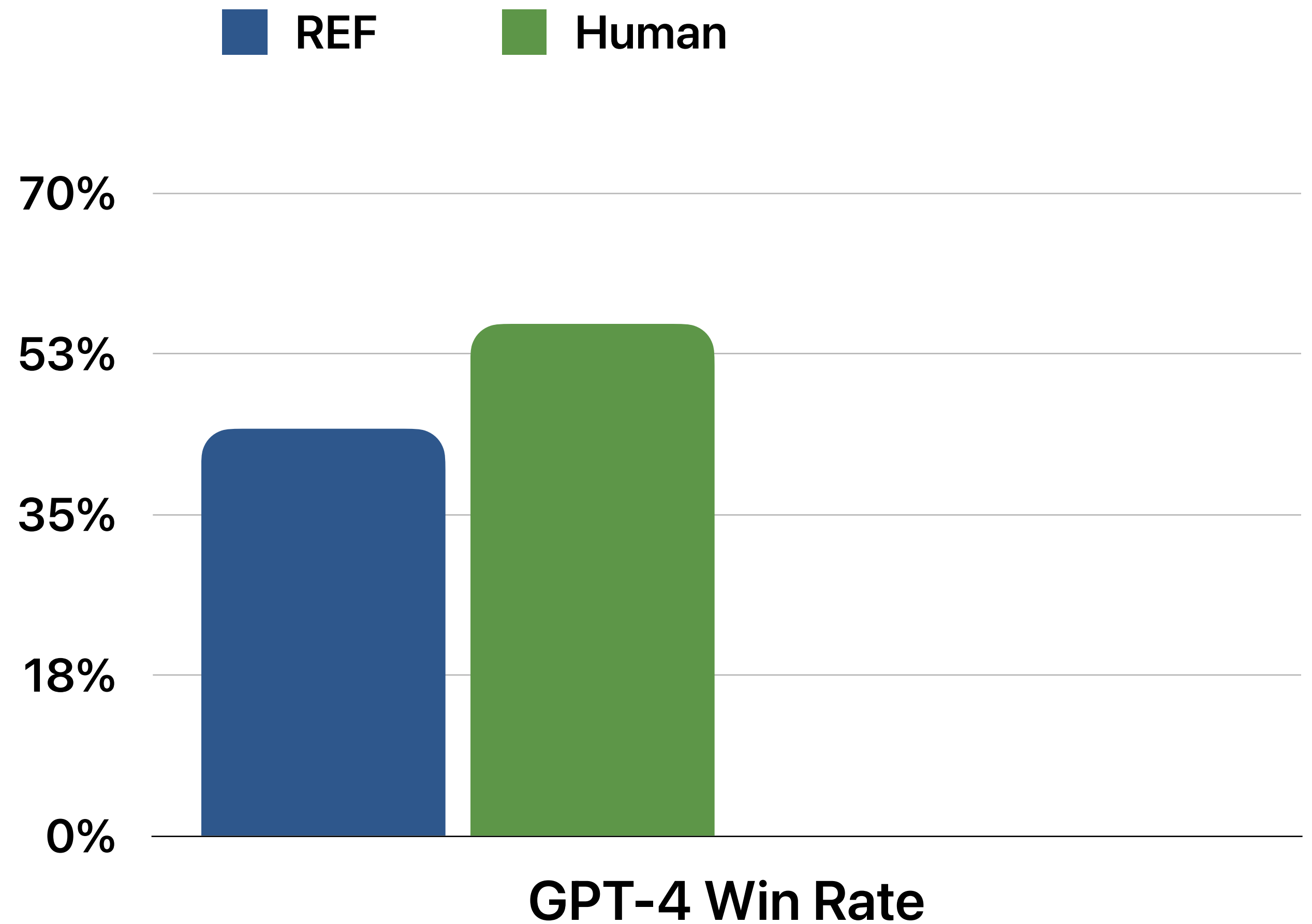
Key Takeaways

- PPO ignores task-specific structure in LLM training
- **Environment resets are a unique property in LLM-based MDPs.**
- PPO++ is a simple algorithm designed to exploit 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?**

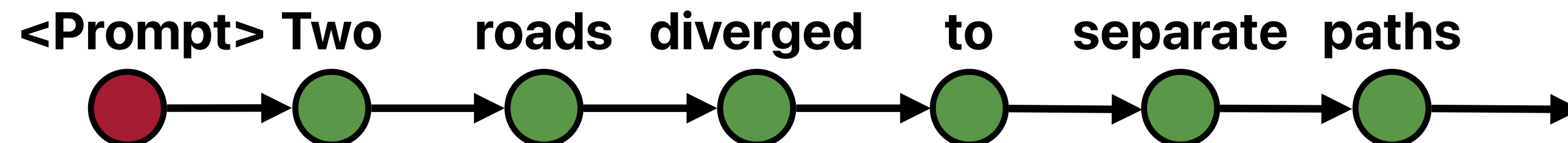
Reset Property

Most text generation tasks
have offline label response

!

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

2. **Reset**



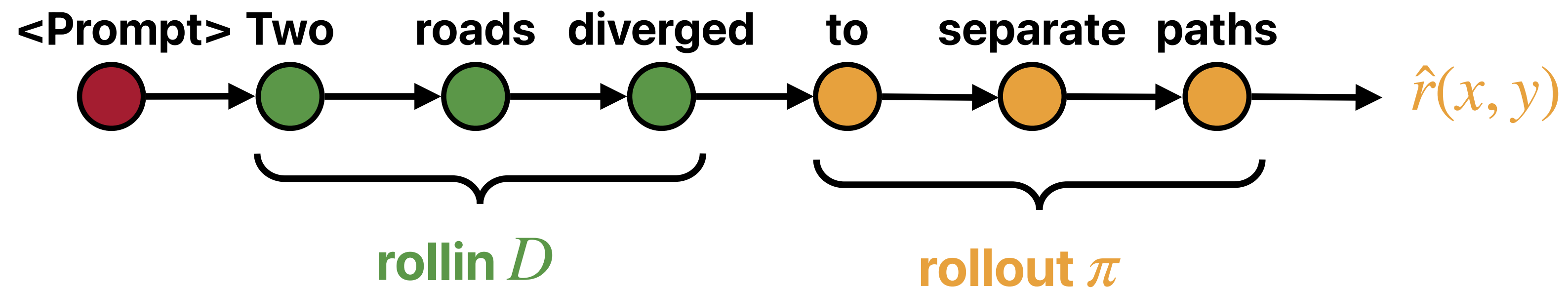
Reset Property

Most text generation tasks
have offline label response

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

2. **Reset**

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



Dataset Reset Policy Optimization



Jonathan D. Chang



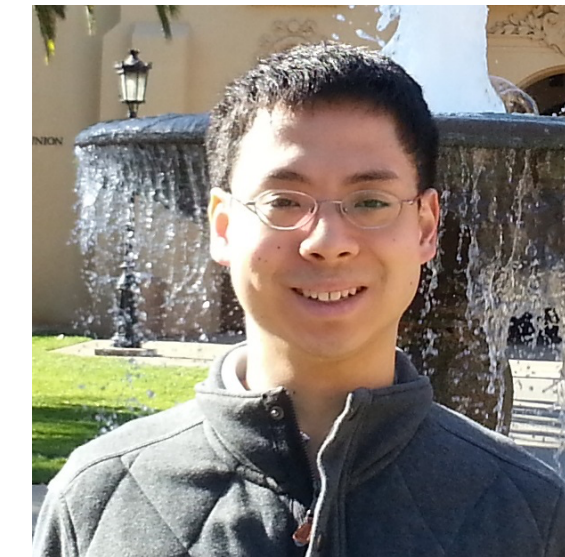
Wenhao Zhan



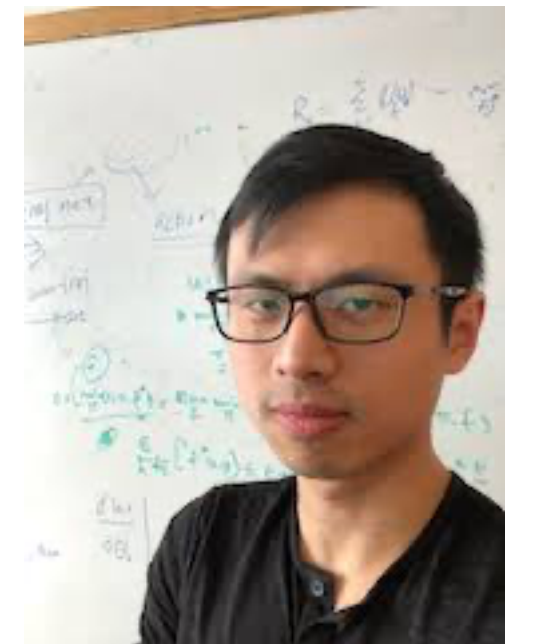
Owen Oertell



Dipendra Misra



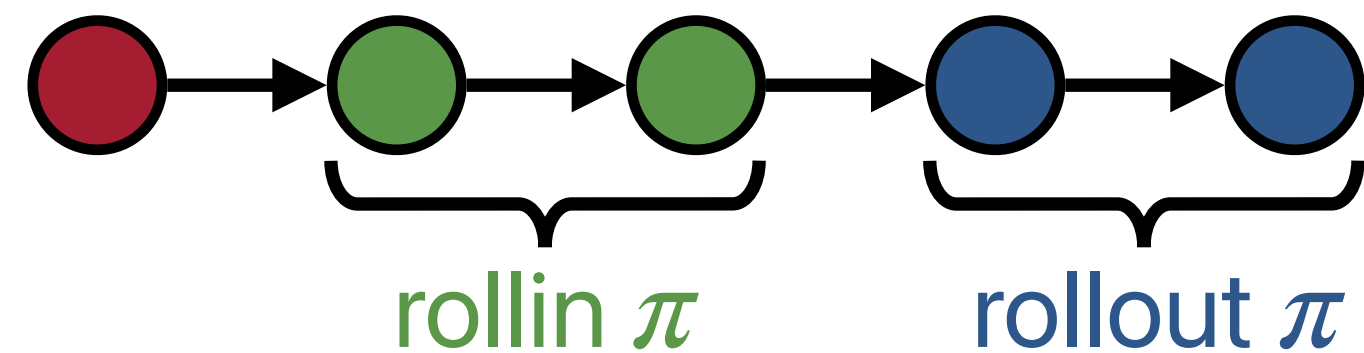
Jason Lee



Wen Sun

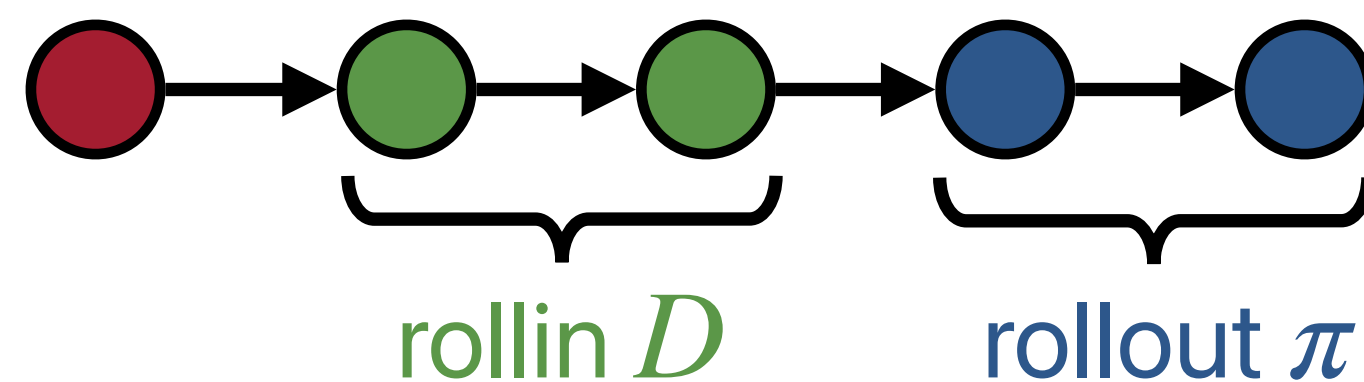
Rollin and Rollout approaches

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

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:

$$\underbrace{\frac{d^{\pi^*}(\tau)}{d^{\pi_{\text{ref}}}(\tau)}}_{\text{Trajectory-wise density}} \leq C_1 < \infty$$

Trajectory-wise density

$$\underbrace{\frac{d^{\pi^*}(x, y)}{d^{\pi_{\text{ref}}}(x, y)}}_{\text{State-action sample-wise density}} \leq C_2 < \infty$$

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.

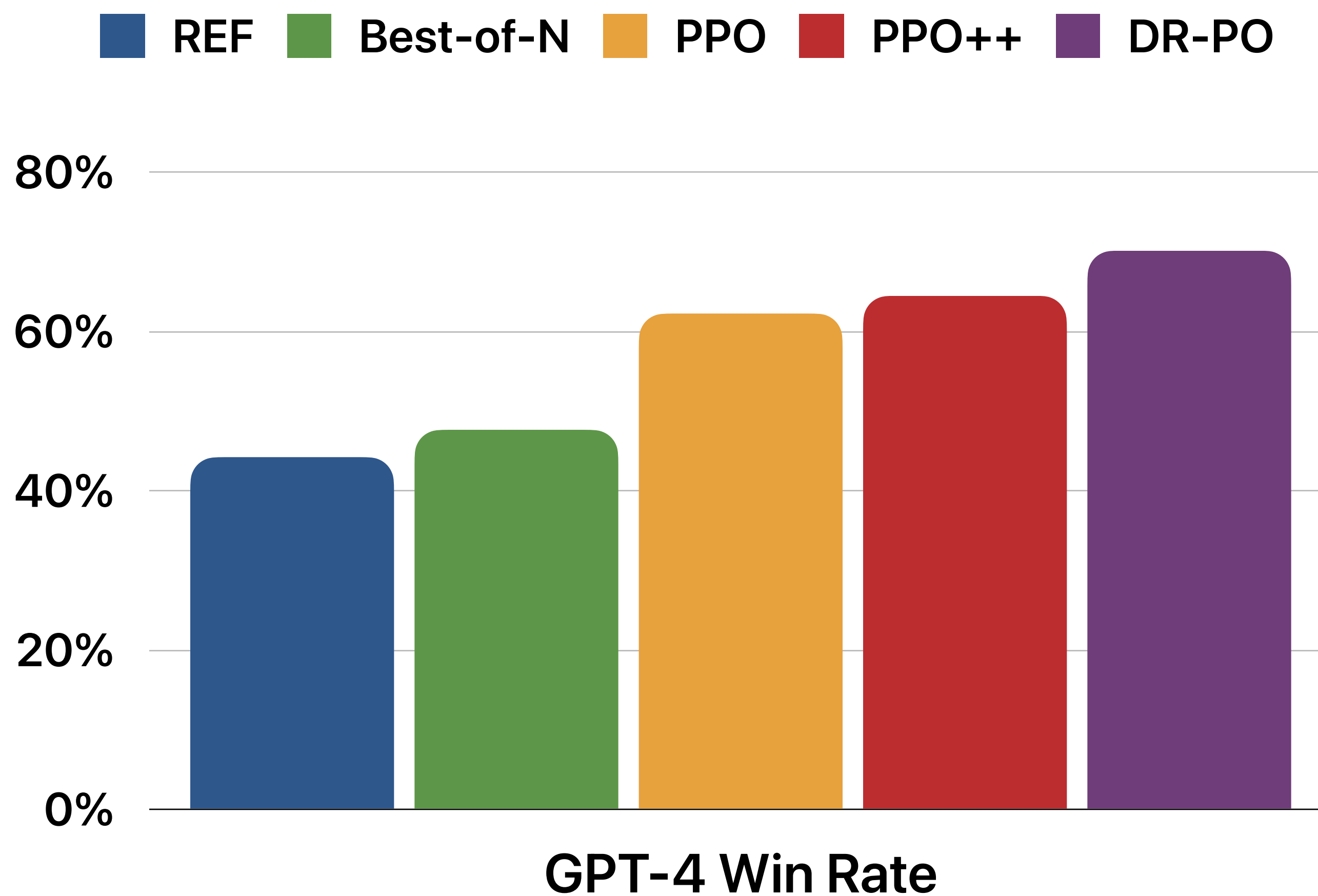
Example Human Label

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

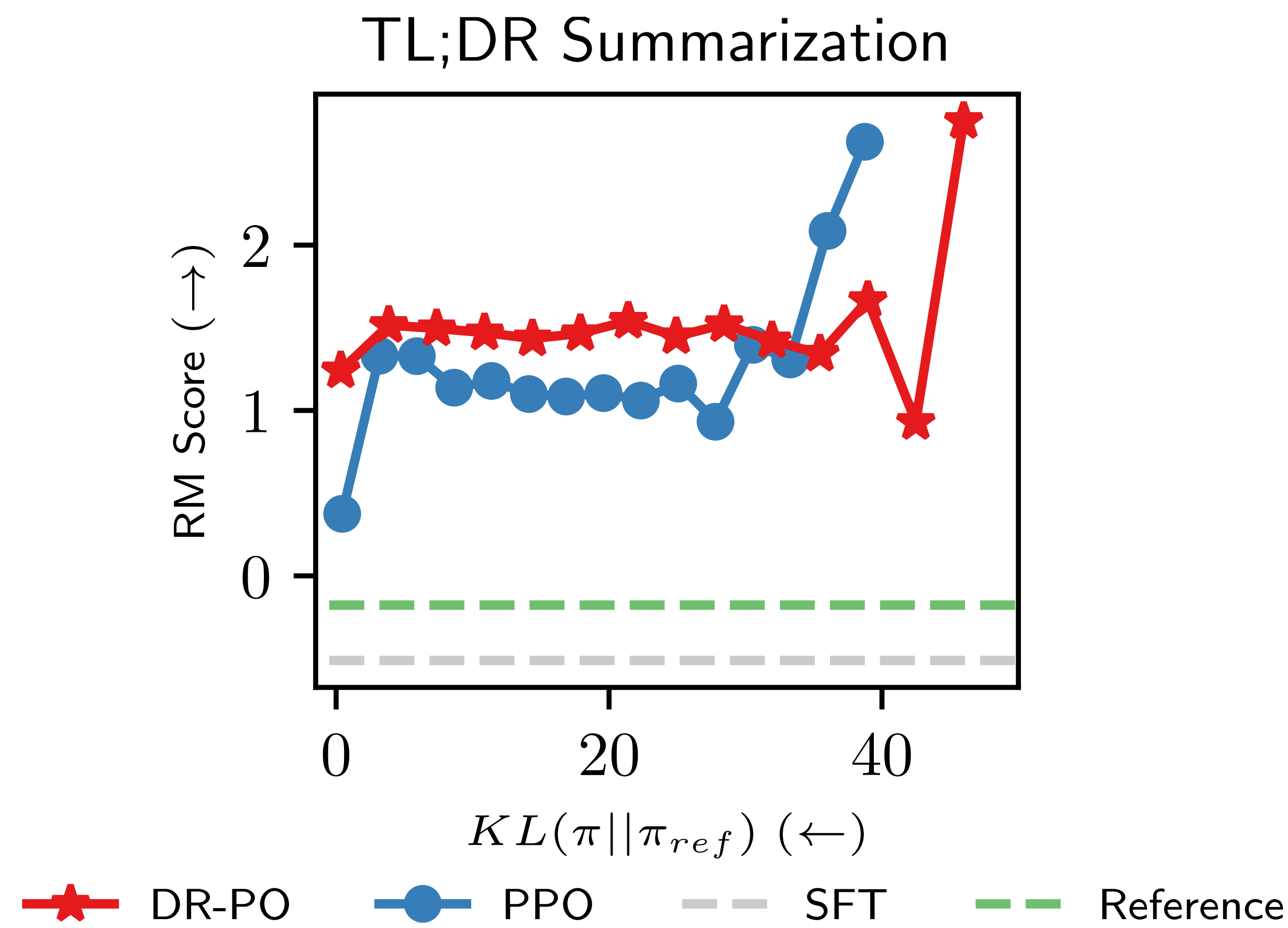
Experimental Results: TL;DR

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.

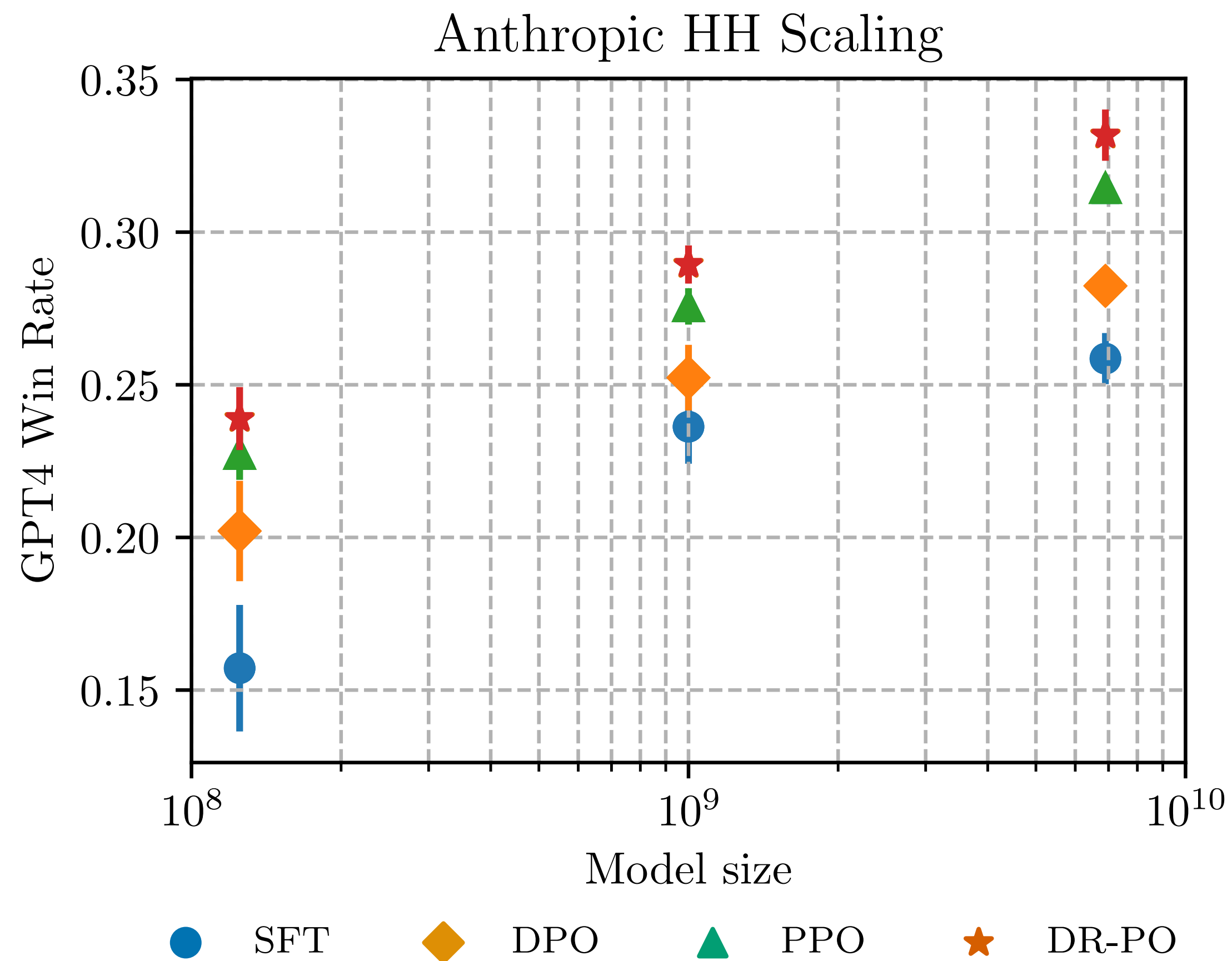
Example Dialogue

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.

Key Takeaways

- Resetting from higher-quality offline demonstrations improves performance.
- DR-PO is provably efficient and improves upon PPO⁺⁺ in theory.
- DR-PO matches PPO in simplicity and computational cost.

Summary

- Resetting DR-PO from offline demonstration data enhances performance.
- DR-PO is provably efficient and improves upon PPO⁺⁺ in theory.
- DR-PO matches PPO in simplicity and **computational cost**.

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 reduce
the compute and memory cost of RL?**

MinXent RL

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

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

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

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

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

$$L(r, D) = - \mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(r(x, y_w) - r(x, y_l) \right) \right] \quad \} \text{ Assume reward follows Bradley Terry}$$

MinXent RL

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

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

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

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

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

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

Rebel: Reinforcement learning via regressing relative rewards

Zhaolin Gao



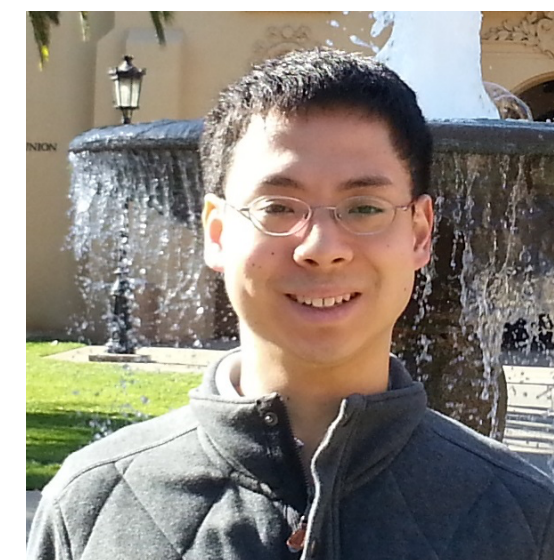
Jonathan D. Chang



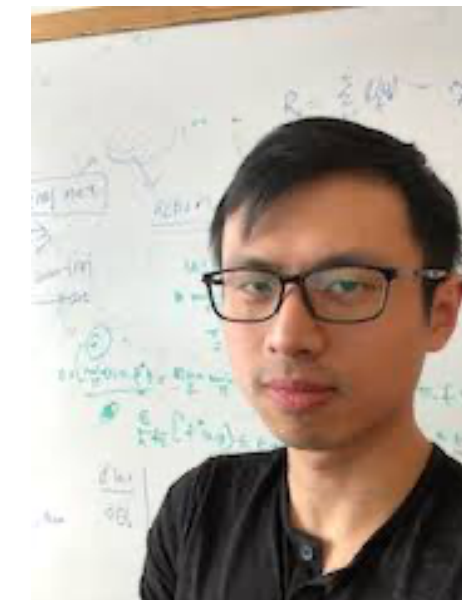
Owen Oertell



Jason Lee



Wen Sun



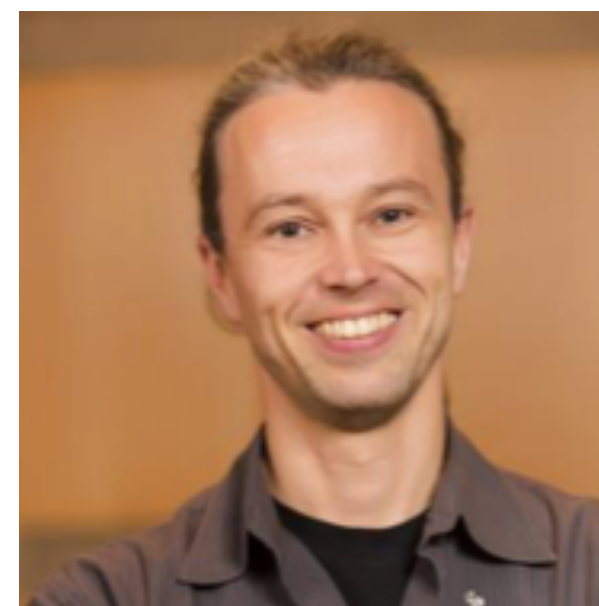
Gokul Swamy



Wenhao Zhan



Thorsten Joachims



J. Andrew Bagnel



MinXent RL

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

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

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

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

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

$$\left(r(x, y) - \frac{1}{\eta} \left(\ln Z(x) + \ln \left(\frac{\pi_{t+1}(y \mid x)}{\pi_t(y \mid x)} \right) \right) \right)^2$$

How can we handle the fact that $Z(x)$ is intractable?

REBEL

algorithm overview

At iteration t with policy π_t

1. Sample (hybrid) data using **resets**:

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

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



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

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

Informal Theory of REBEL

Informal statement:

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

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

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

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

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

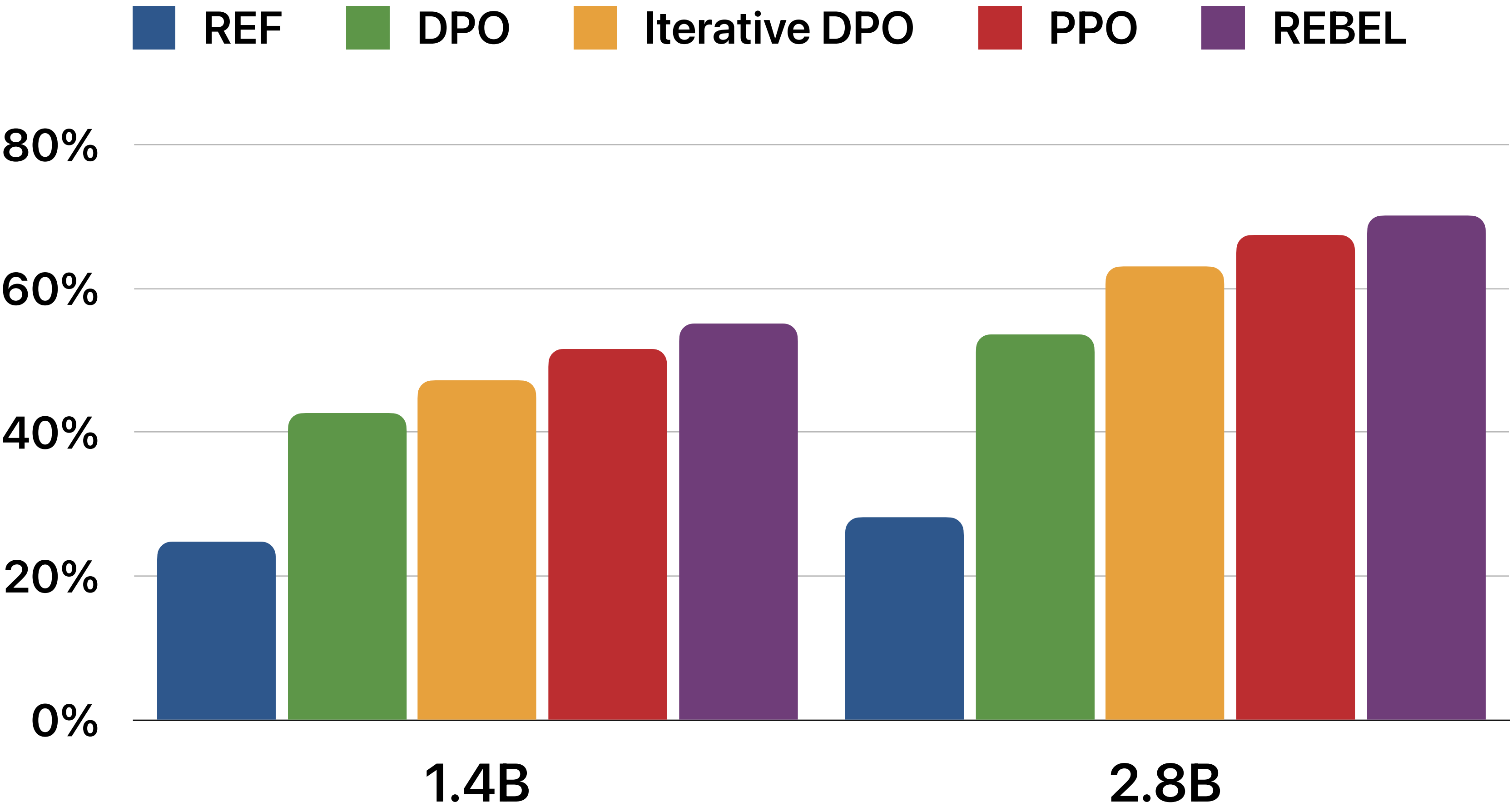
POST: I have a three year old Dober/Pit mix named Romulus ("Rome" for short). I live with 3 other dogs: a 10 year old labrador, a 2 year old French Bulldog and a 8 year old maltese mix. The four of them get along just fine, Rome and the Frenchie are best best best best friends. He isn't the best at meeting 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

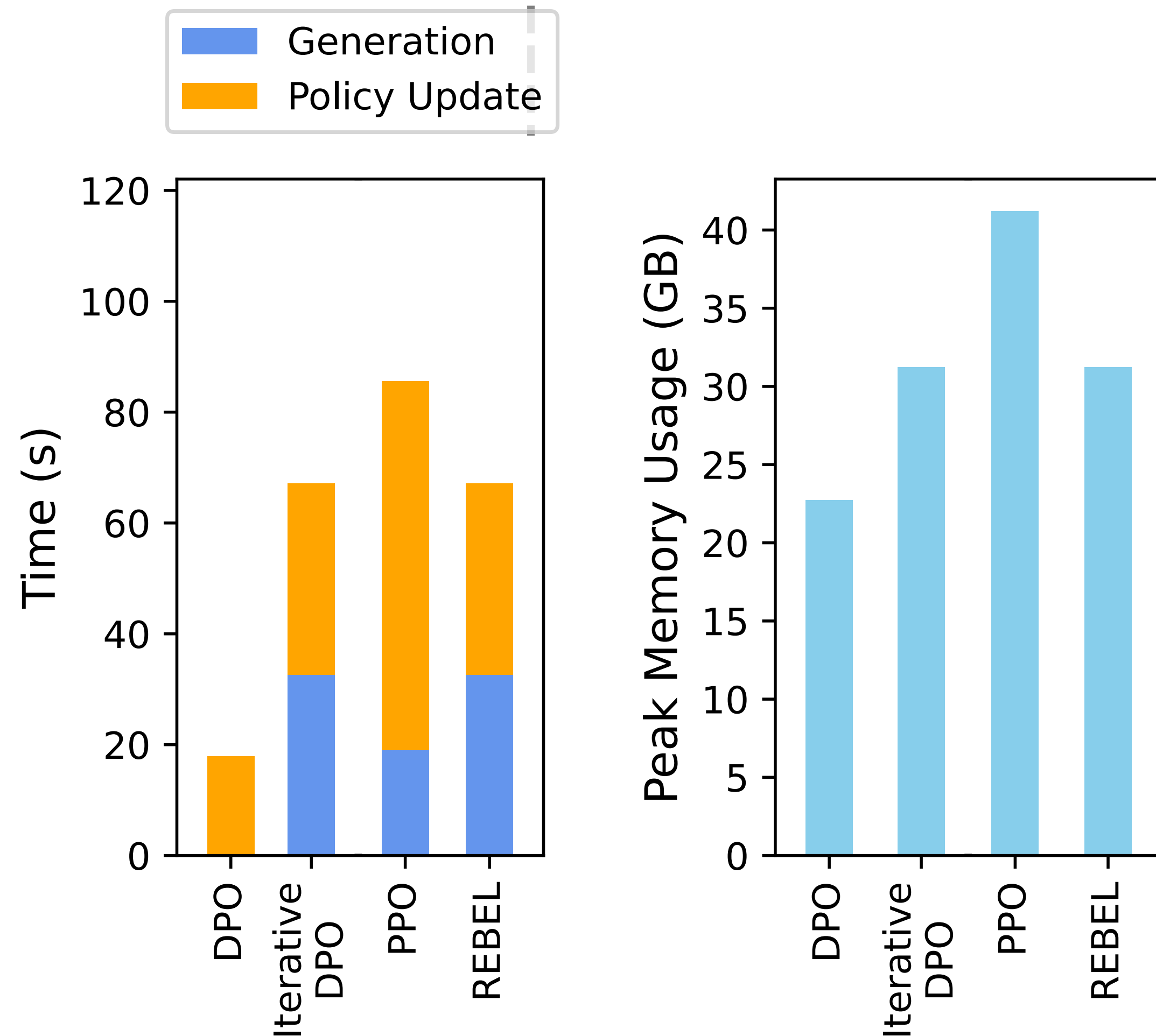
TL;DR: HOW do I introduce new people? HOW do I introduce new dogs? WHAT do I do about 4th of July??

Experimental Results: TL;DR

- Takeaways:**
- 1. Online methods outperform offline methods
 - 2. REBEL consistently outperforms all baseline methods across scales.



Experimental Results



Takeaways:

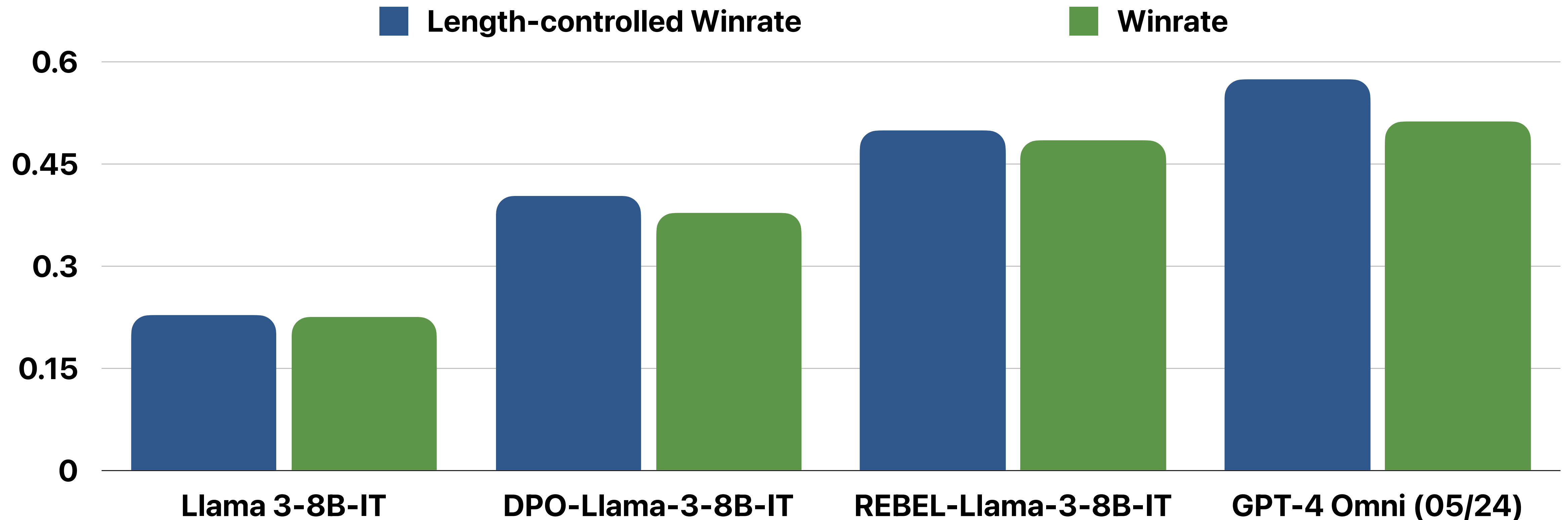
1. **Offline methods** are more efficient but yield lower win rates.
2. REBEL outperforms PPO in both win rate and resource efficiency.
3. REBEL outperforms iterative DPO in win rate while matching its efficiency.

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 \[Wang et al\]](#)

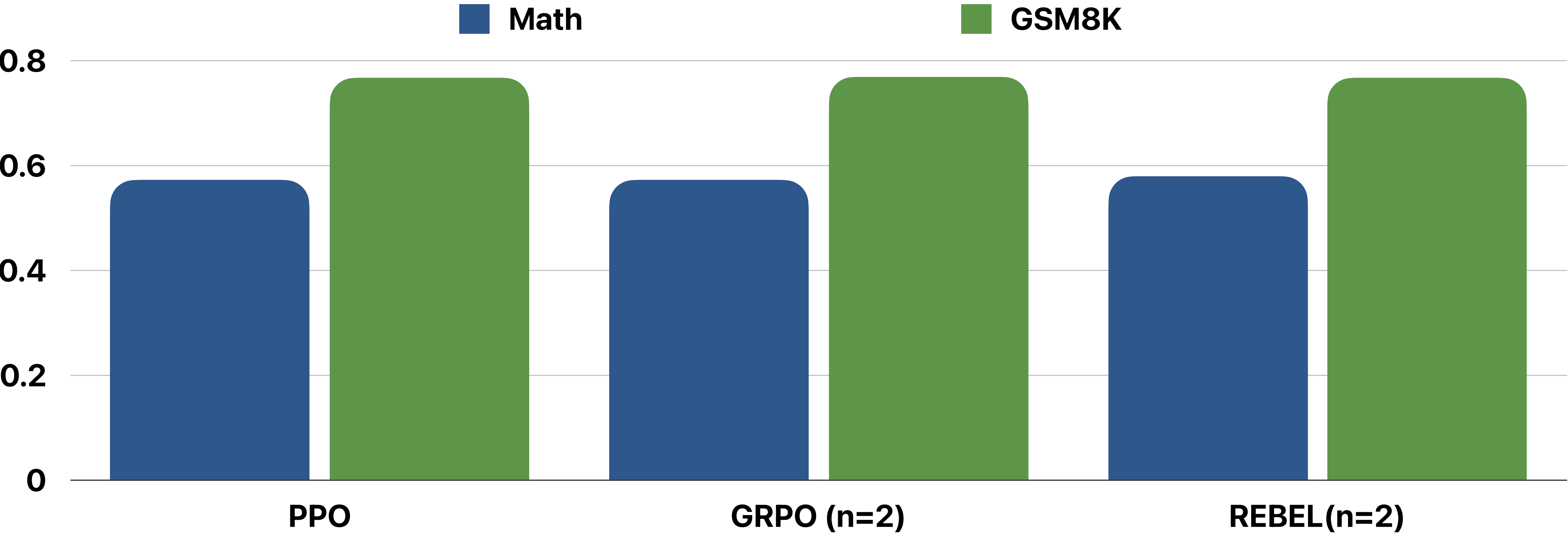


Scaling to Reasoning Tasks

Experimental Results

Fine-tuning Qwen-2.5B model

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



Key Takeaways

- REBEL reframes RL as a sequence of relative reward regression problems.
- Empirically, REBEL is faster, more memory-efficient, and performs better than PPO.
- REBEL achieves strong results on standard LLM benchmarks.

How can environment resets be utilized to solve the MinxEnt RL objective efficiently?

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

Acknowledgement

