Efficient Policy Optimization Techniques for LLMs

Kianté Brantley

Harvard John A. Paulson
School of Engineering
and Applied Sciences

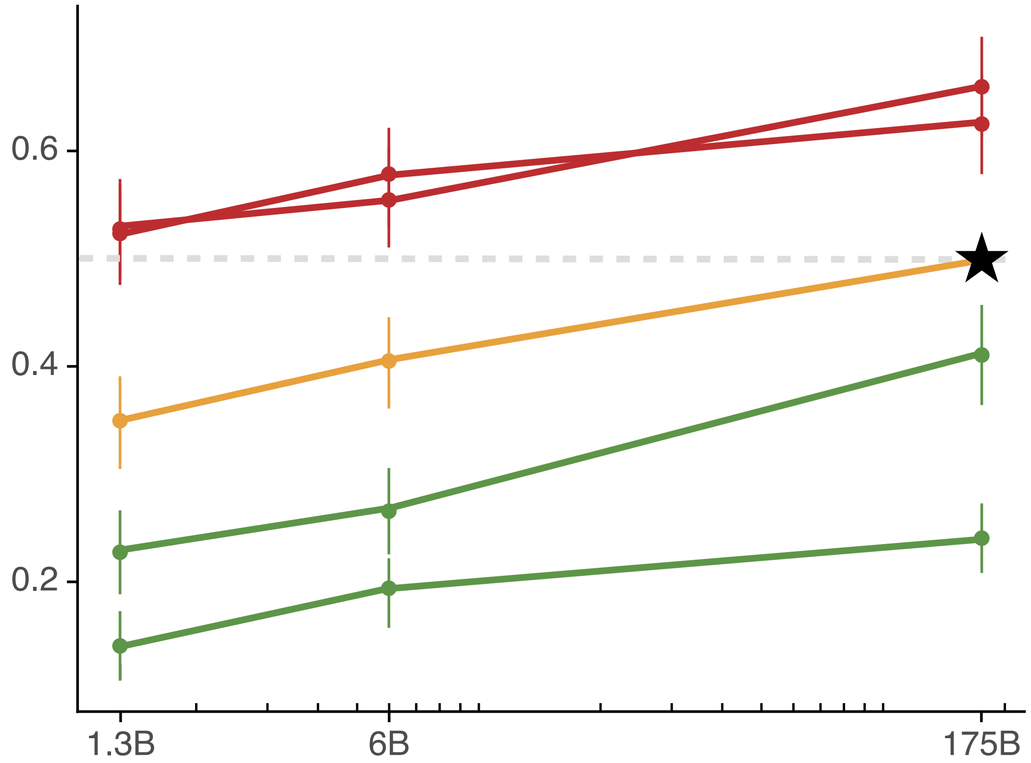
“A large language model is a type of deep learning model that is trained on massive text datasets to understand and generate human language.”

Large Language Model

Learning from Human Feedback

Pre-Training | Supervised Fine-Tuning | RL Fine-Tuning

Human Win rate against reference model 175B



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

Learning from Human Feedback

Step 1: Supervised Fine Tuning (SFT)

Given: $D = \{(\textbf{Prompts, Desired Generations})\}$

Optimize: Negative Log Likelihood of generations in D

Step 2: Preference Reward Model Training

Given: $D = \{(\textbf{Prompts, Gen 1, Gen 2,..., Gen K})\}$

Optimize: Pairwise ranking loss of ordered generations

Learning from Human Feedback

Step 1: Supervised Fine Tuning (SFT)

Given: $D = \{(\textbf{Prompts, Desired Generations})\}$

Optimize: Negative Log Likelihood of generations in D

Step 2: Preference Reward Model Training

Given: $D = \{(\textbf{Prompts, Gen 1, Gen 2,..., Gen K})\}$

Optimize: Pairwise ranking loss of ordered generations

?

Step 3: Reinforcement Learning

Given: LLM from **(step 1)** π_{ref}

Optimize: Reward Model from **(step 2)** $\hat{r}(x, y)$

RL Algorithm Used: Proximal Policy Optimization (PPO)

$$\arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{\substack{x \sim D \\ y \sim \pi(\cdot \mid x)}} [\hat{r}(x, y)] - \frac{1}{\eta} D_{\text{KL}}\left(\pi \| \pi_{\text{ref}}\right)$$

How to design efficient algorithms that can solve this KL-Regularized RL objective?

Outline

- **Reset with reference policy**
- **Reset with the demonstration data**
- **Regress the relative rewards**
- **Regress the relative future rewards**

Outline

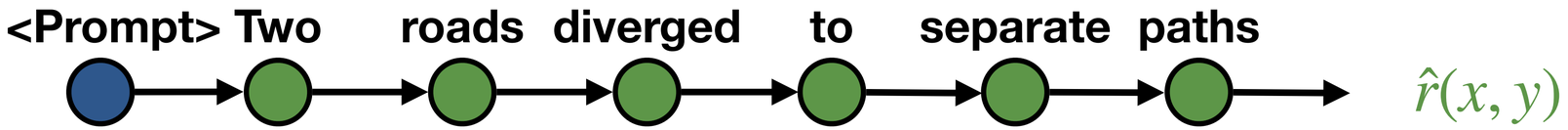
- **Reset with reference policy**

Reset Property

Reset allows us to rollout a policy from partial sentences ?

Inject other sources of data in experience collection!

1. Sample a prompt from D
2. Sample a generation from π



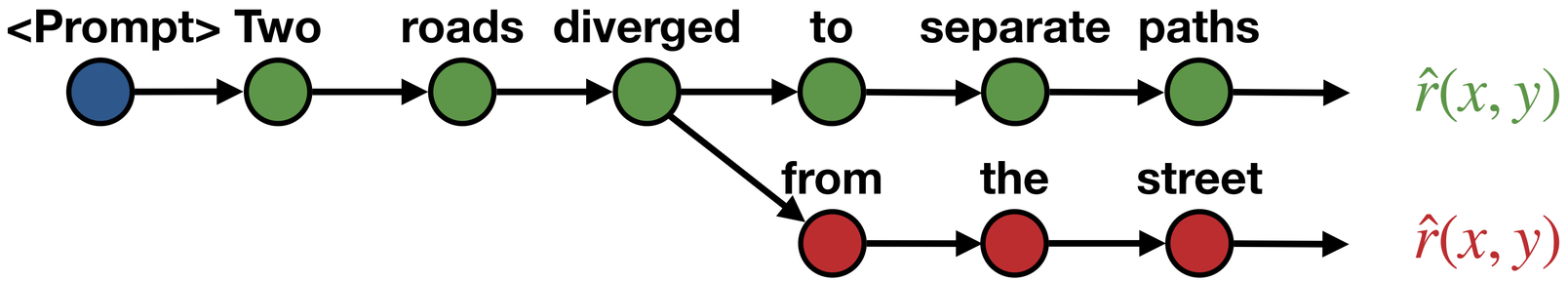
Reset Property

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

Reset allows us to rollout a policy from partial sentences ?

Inject other sources of data in experience collection!

1. Sample a prompt from D
2. Sample a generation from π
3. Reset and complete the generation from the partial sequence



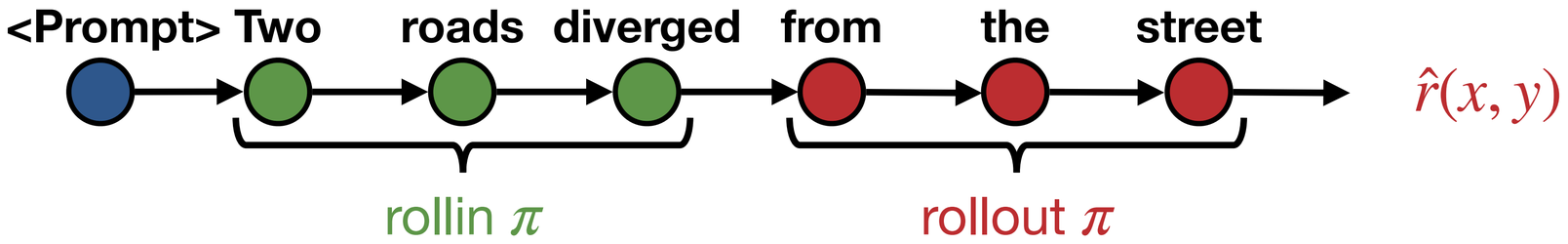
Reset Property

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

Reset allows us to rollout a policy from partial sentences ?

Inject other sources of data in experience collection!

1. Sample a **prompt** from D
2. Sample a **generation** from π
3. **Reset** and complete the **generation** from the partial sequence

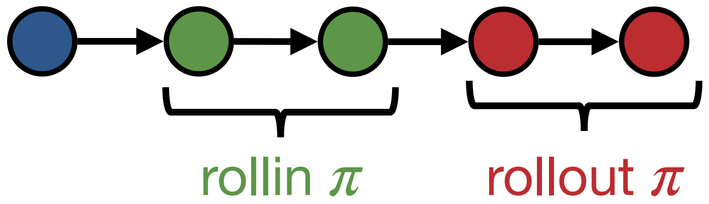


Rollin and Rollout
approaches

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

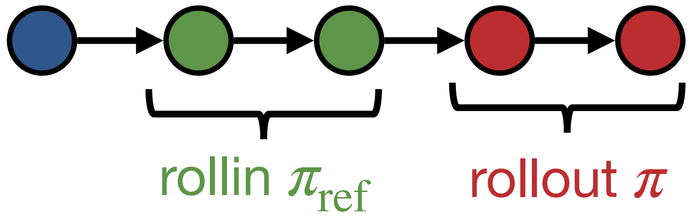
We want to remain close to π_{ref}

PPO (RL algorithm)



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

PPO^{++} [CBRMS Instruction Workshop 2023]



- **Samples prompts** $x \sim \beta D + (1-\beta) d^{\pi_{\mathrm{ref}}}$
- Scores actions with $\hat{r}(x,y)$
- **Intuition: Richer initial state distribution**

Proximal Policy Optimization, John Schulman et al., 2017
LEARNING TO GENERATE BETTER THAN YOUR LLM, Jonathan Chang et al., 2023

Theory of PPO++

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

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

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

Performance gap

$$J(\pi^\star) - J(\pi^t) \leq O\left(H^2 \max_s \left(\frac{d^{\pi^\star}(s)}{d^{\pi_{\text{ref}}}(s)}\right)\epsilon\right)$$

Assume bound density ratio and π_{ref} provides coverage for $\pi^\star$

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.
TL;DR:

Example Human Label

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

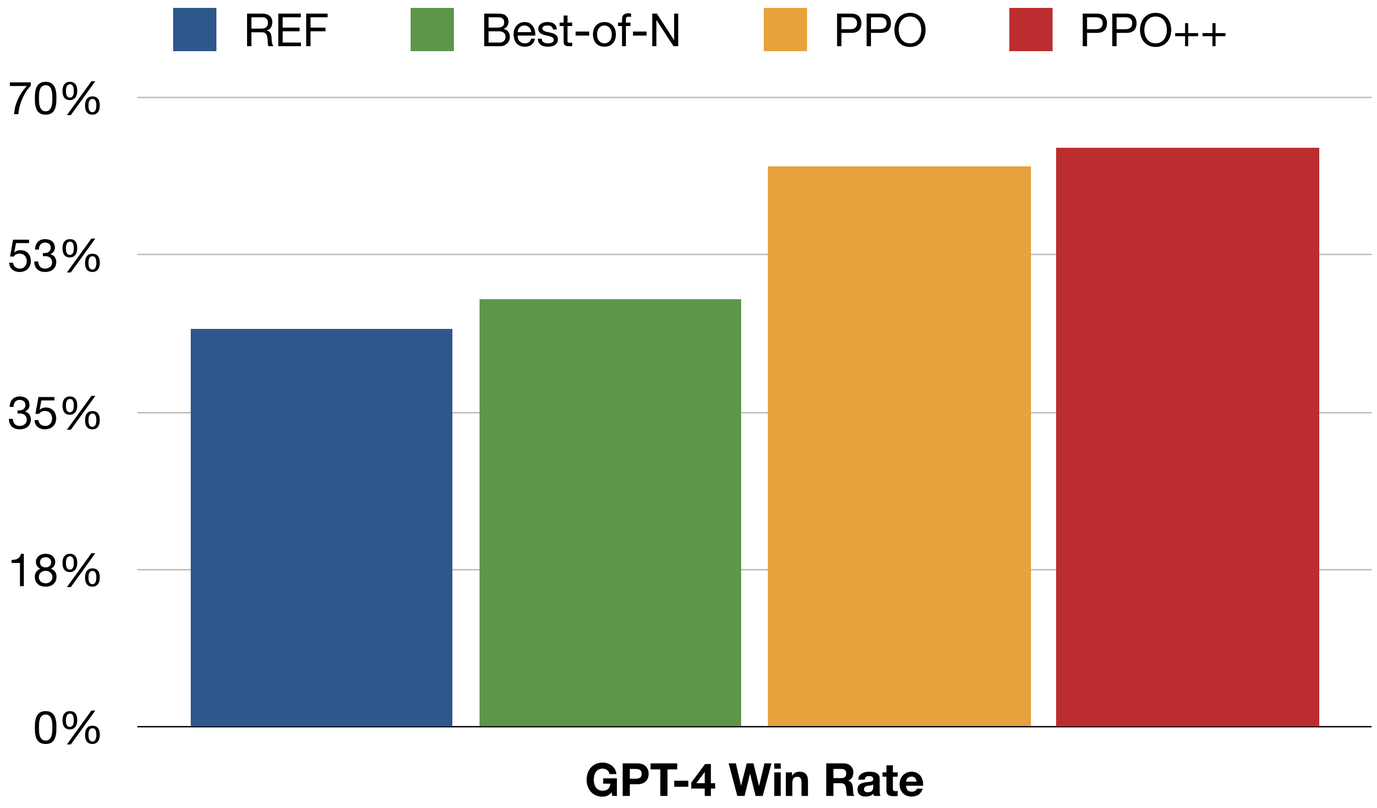
Learning to summarization from human feedback Stiennon et al. 2017

Experimental Results: TL;DR Summarization

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>

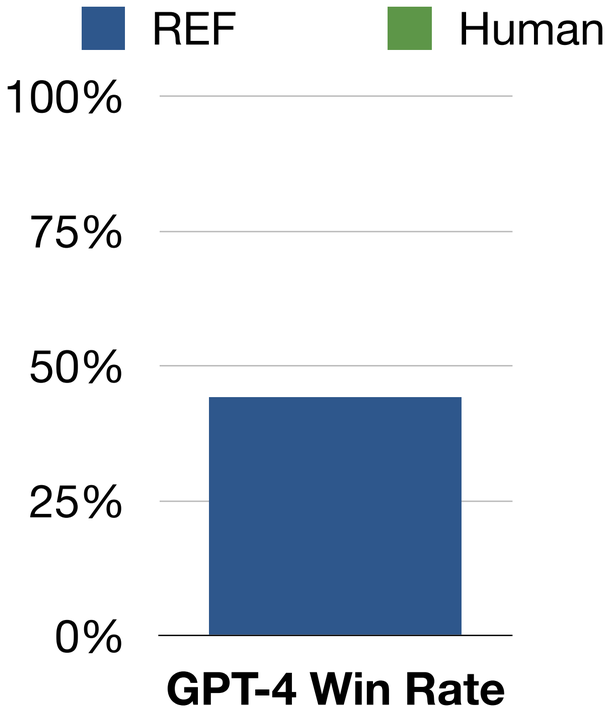


Summary

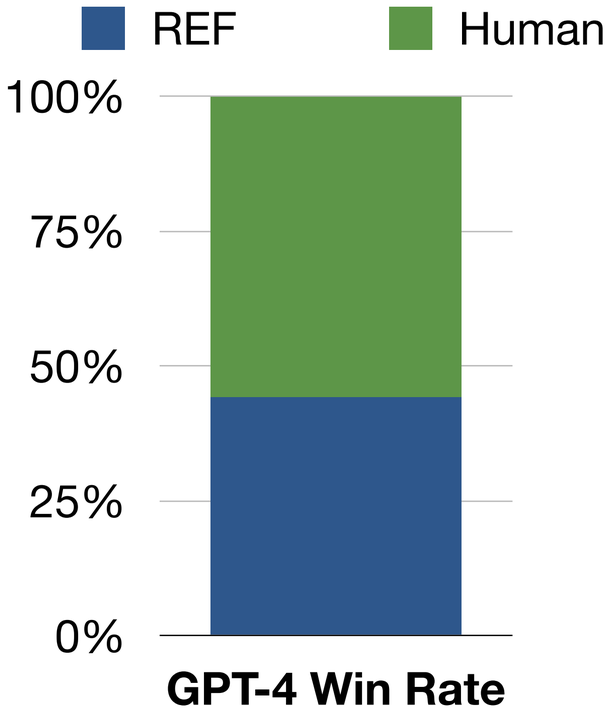
- PPO Does not utilize the specific structure of the problem.
- **The ability to reset** is a special property of MDPs for LLMs
- PPO^{++} is a simple algorithm that uses the ability to reset.

Outline

- **Reset with reference policy**
- **[Now]: Reset with the demonstration data**



What is a better mixing distribution to enhance our initial state distribution?

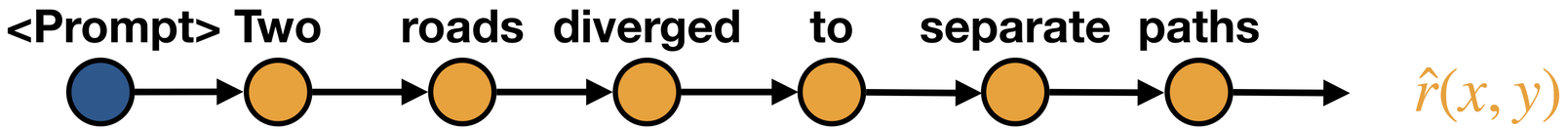


What is a better mixing distribution to enhance our initial state distribution?

Dataset Reset Policy Optimization (DR-PO)

Most text generation tasks have offline label generations

1. Sample a prompt and generation from D

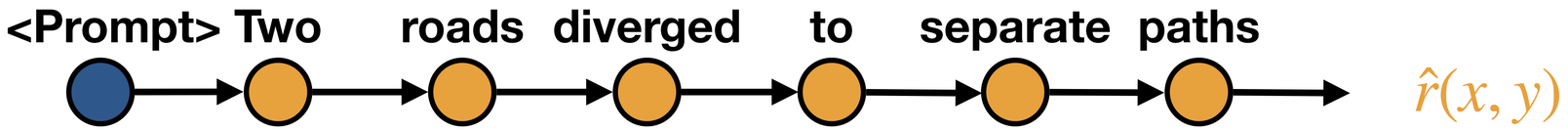


Dataset Reset Policy Optimization for RLHF, Jonathan Chang et al., 2024

Dataset Reset Policy Optimization (DR-PO)

Most text generation tasks have offline label generations

1. Sample a prompt and generation from D
2. Reset and sample a generation from π

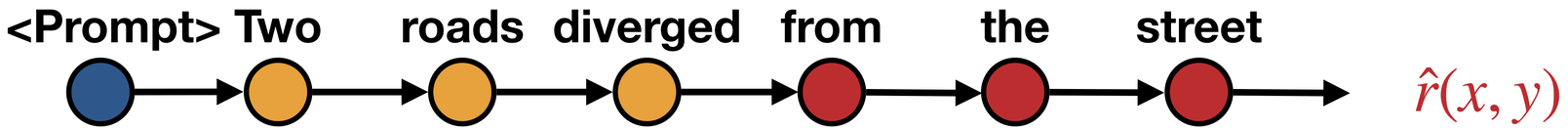


Dataset Reset Policy Optimization for RLHF, Jonathan Chang et al., 2024

Dataset Reset Policy Optimization (DR-PO)

Most text generation tasks have offline label generations

1. Sample a prompt and generation from D
2. Reset and sample a generation from π



Dataset Reset Policy Optimization for RLHF, Jonathan Chang et al., 2024

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:

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

Trajectory-wise density

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

State-action sample-wise density

Experimental Setup

Task Statement

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

Example Post

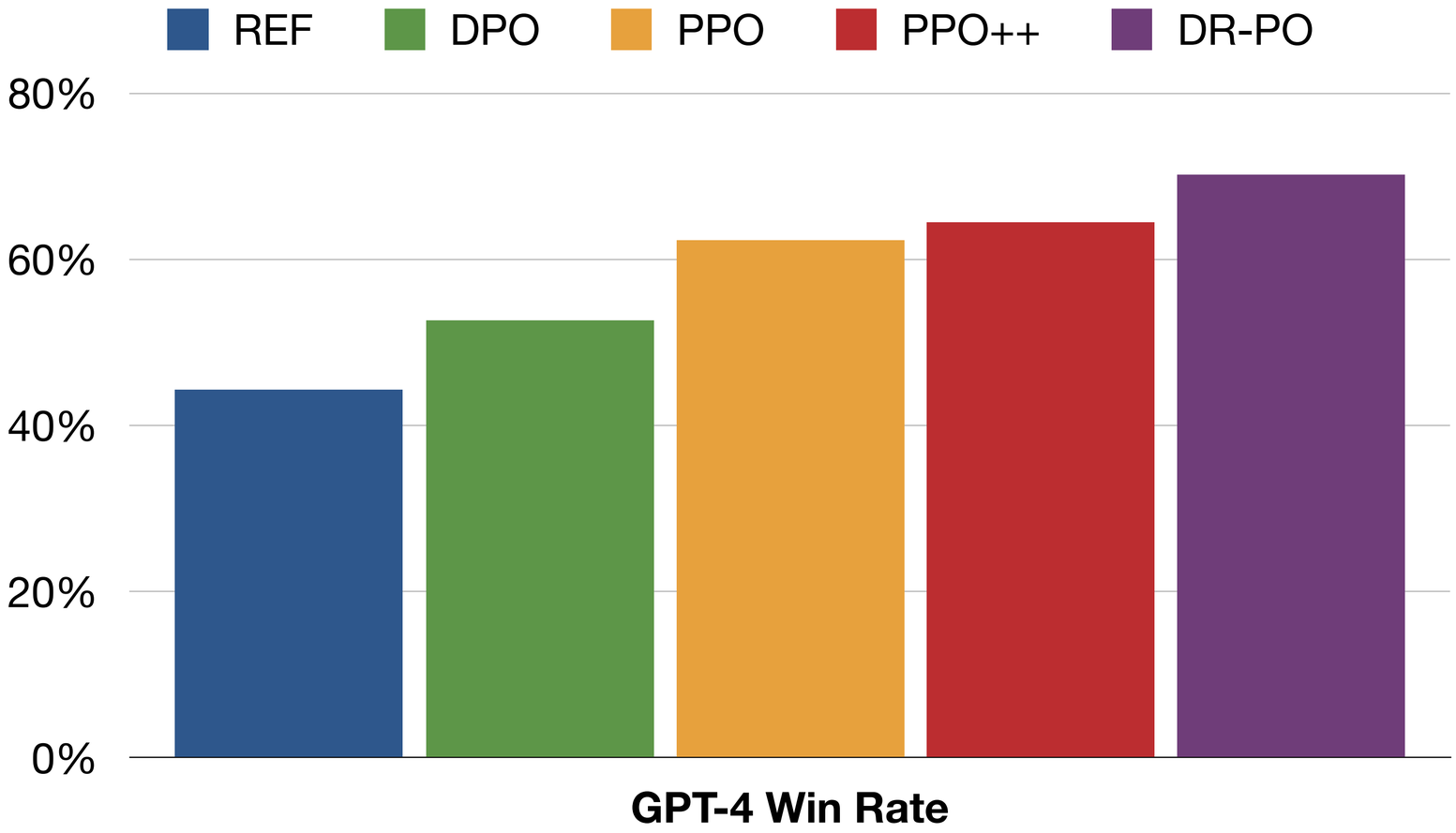
SUBREDDIT: r/dogs
TITLE: [HELP] Not sure how to deal with new people/dogs and my big ole pup
POST: I have a three year old Dober/Pit mix named Romulus ("Rome" for short). I live with 3 other dogs: a 10 year old labrador, a 2 year old French Bulldog and a 8 year old maltese mix. The four of them get along 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.
TL;DR:

Example Human Label

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

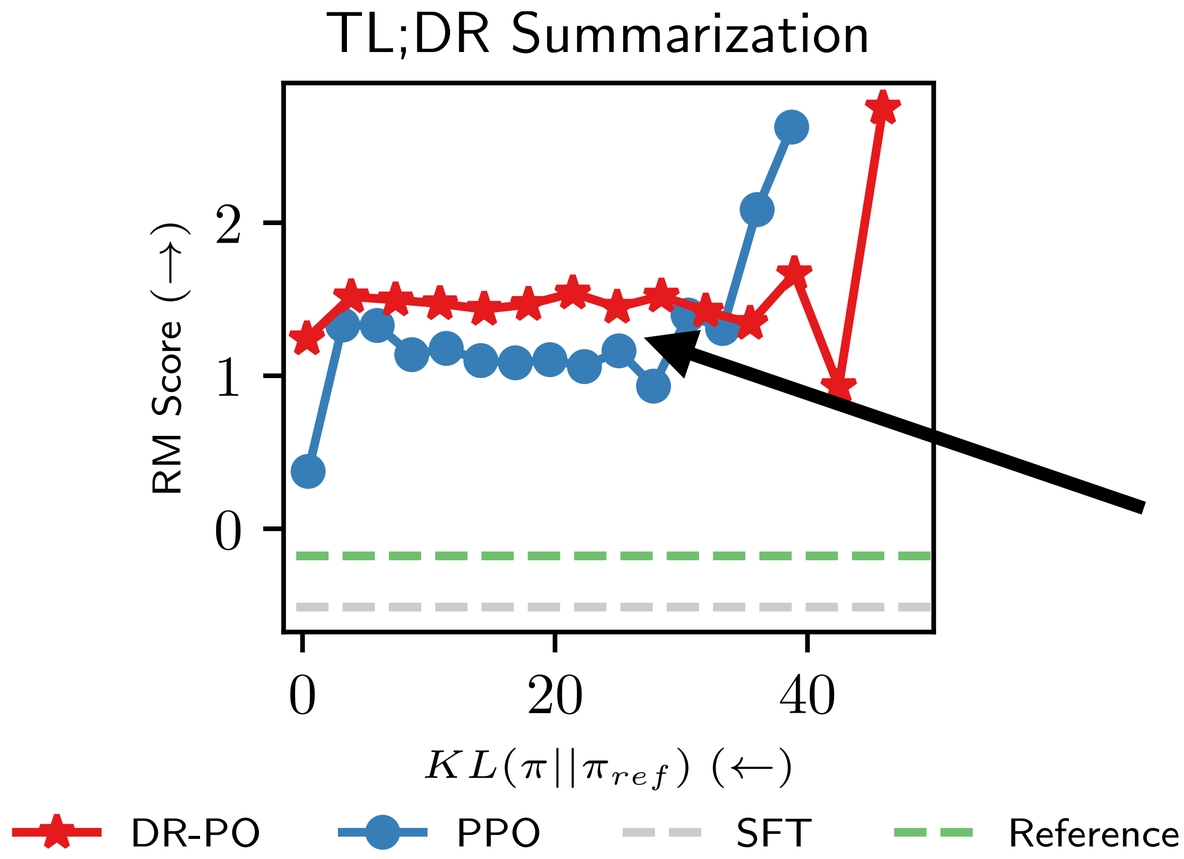
Experimental Results: TL;DR Summarization



Takeaways:

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

Experimental Results: TL;DR



Takeaways:

DR-PO achieves a higher RM score with lower KL across most reward values

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

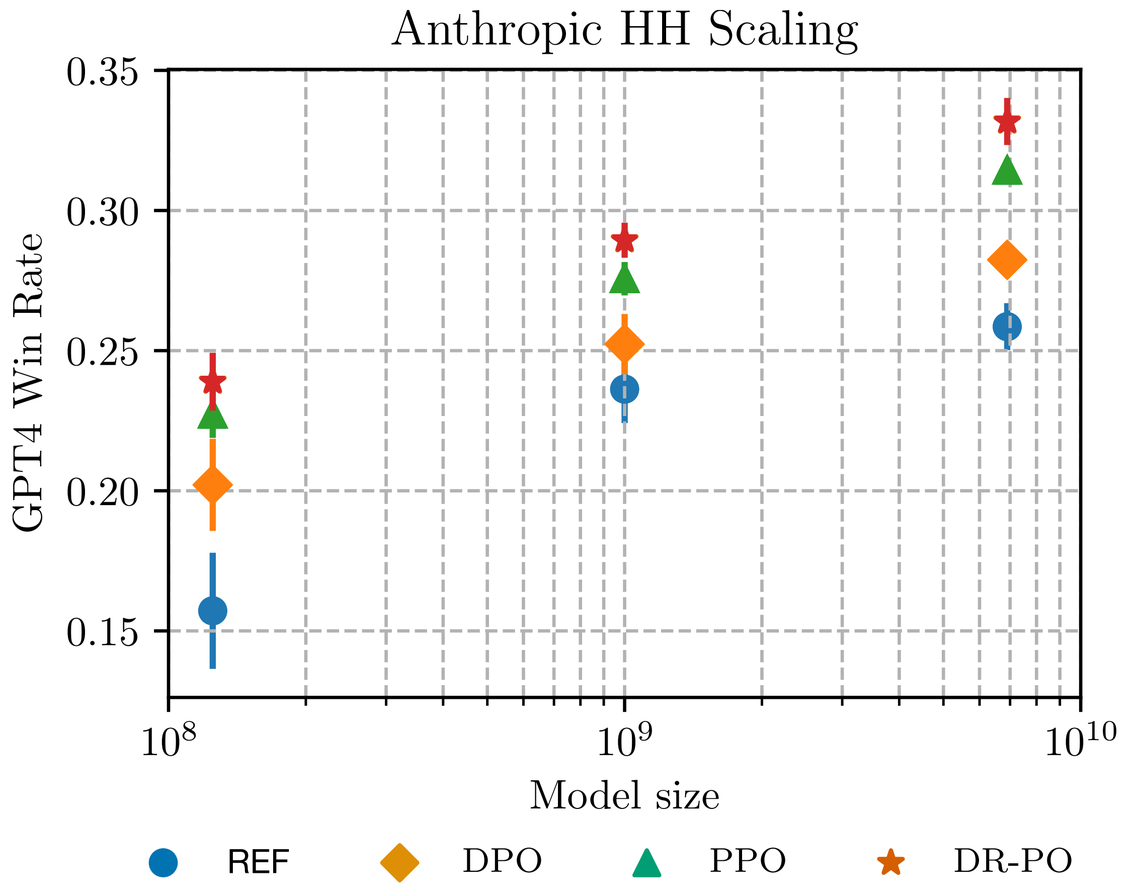
Human: What do I do if I crack a molar?

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.

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, Yuntao Bai et al., 2022

Experimental Results



Takeaways:

1. Online algorithms perform better than offline algorithms.
2. DR-PO outperforms all baseline algorithms across every scale.

Summary

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

Outline

- **Reset with reference policy**
- **Reset with the demonstration data**
- **[Now]: Regress the relative reward**

Summary

- Resetting DR-PO directly from offline demonstration data enhances performance.
- DR-PO is provably efficient and improves upon PPO^{++} in theory
- DR-PO is just as straightforward as PPO, requiring no additional computation or less computation.

How can we reduce the computation and memory overhead of current RL algorithms?

KL-Regularized RL

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

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

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

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

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

KL-Regularized RL

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

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

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

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

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

$$\left(\frac{1}{\eta} \left(\ln \frac{\pi(y \,|\, x)}{\pi_t(y \,|\, x)} - \ln \frac{\pi(y' \,|\, x)}{\pi_t(y' \,|\, x)} \right) - \left(r(x, y) - r(x, y') \right) \right)^2$$

Regress the difference in rewards to cancel $Z(x)$:

Regressing Relative Reward Based RL (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:

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

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

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

Informal Theory of REBEL

Informally

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

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

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

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

Experimental Setup

Task Statement

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

Example Post

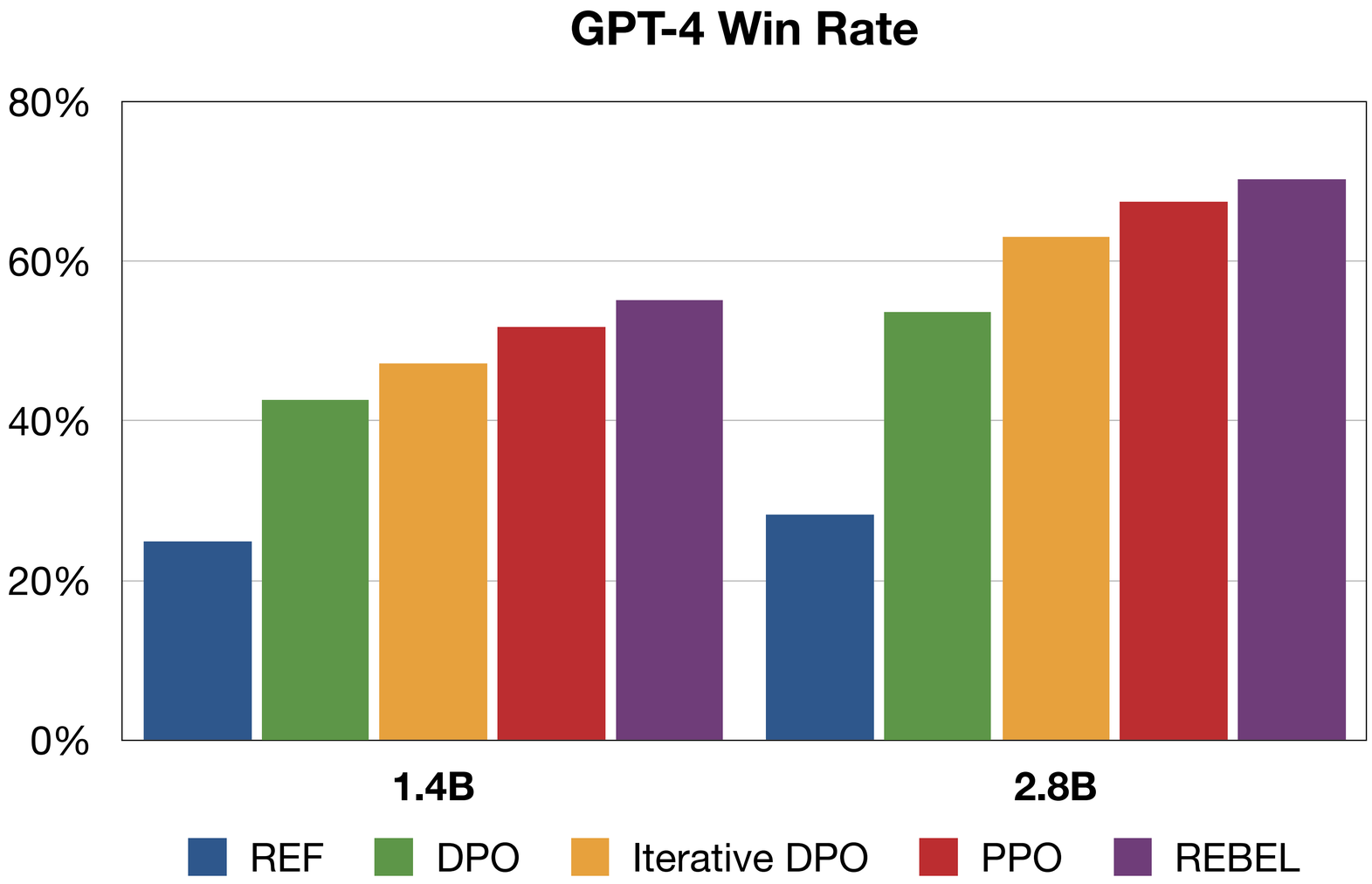
SUBREDDIT: r/dogs
TITLE: [HELP] Not sure how to deal with new people/dogs and my big ole pup
POST: I have a three year old Dober/Pit mix named Romulus ("Rome" for short). I live with 3 other dogs: a 10 year old labrador, a 2 year old French Bulldog and a 8 year old maltese mix. The four of them get along 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.
TL;DR:

Example Human Label

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

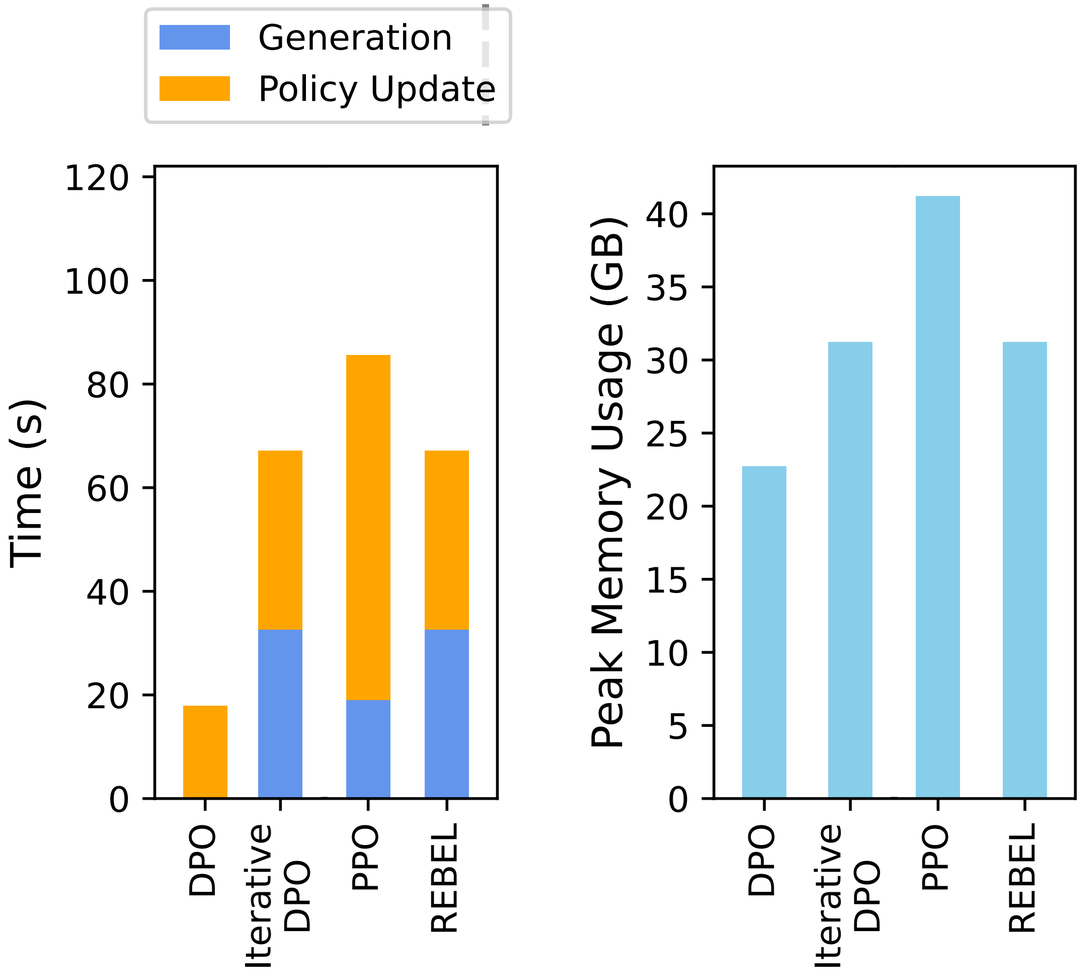
Experimental Results



Takeaways:

1. RL can do better than humans
2. Online methods outperform pure offline method DPO
3. REBEL outperforms PPO

Experimental Results



Takeaways:

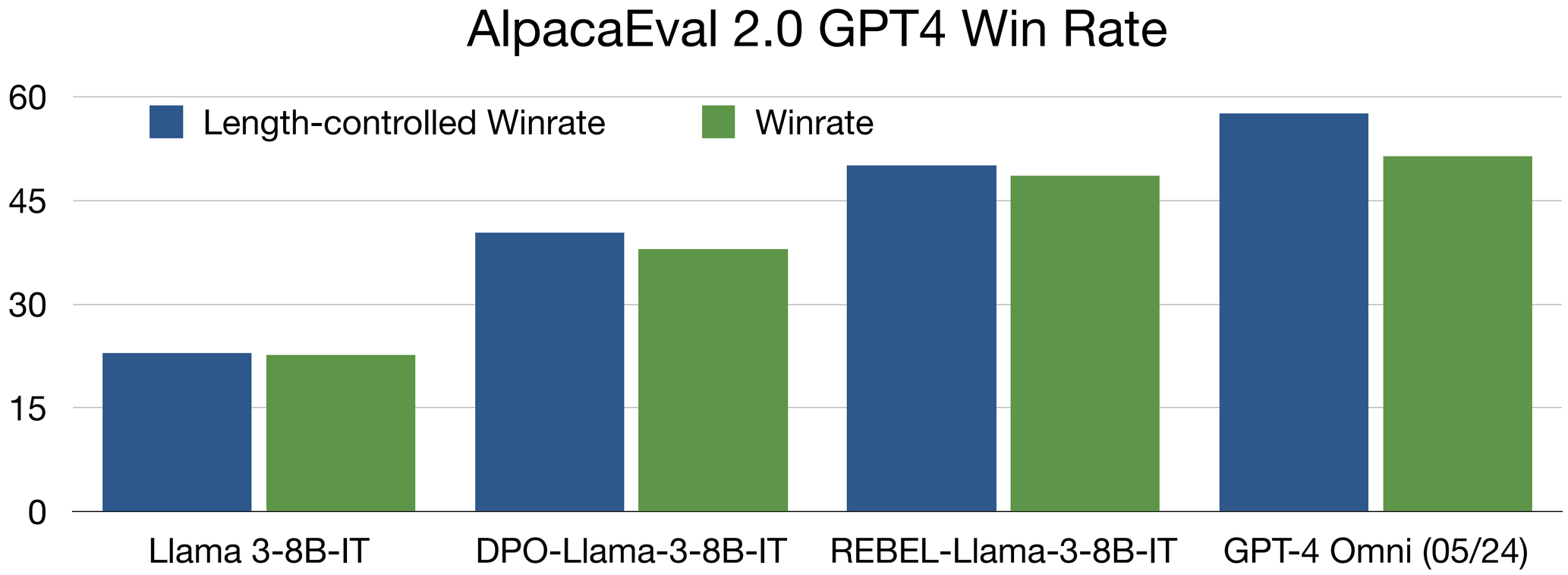
1. Offline methods take less time and memory, but they result in lower win rates.
2. REBEL performs better than PPO in both win-rate and efficiency in computation and memory usage.

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]



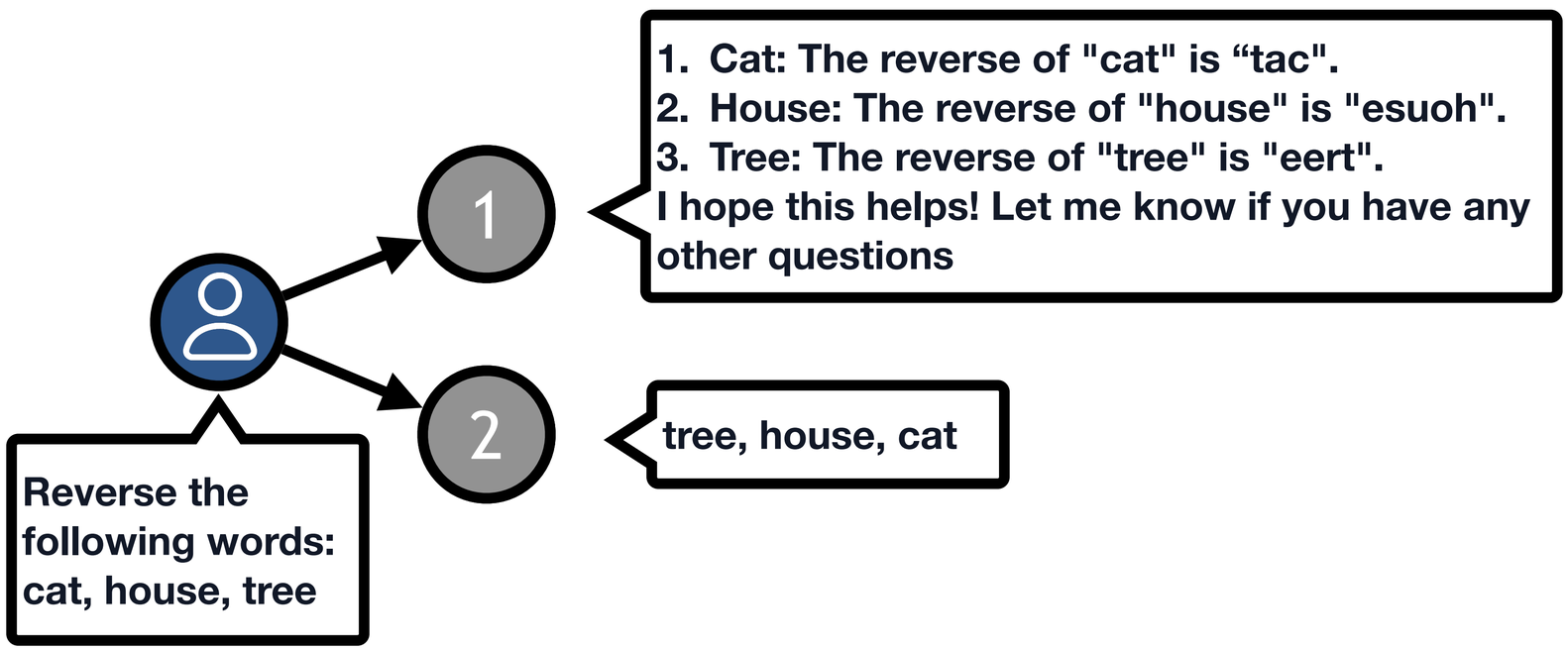
Summary

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

Outline

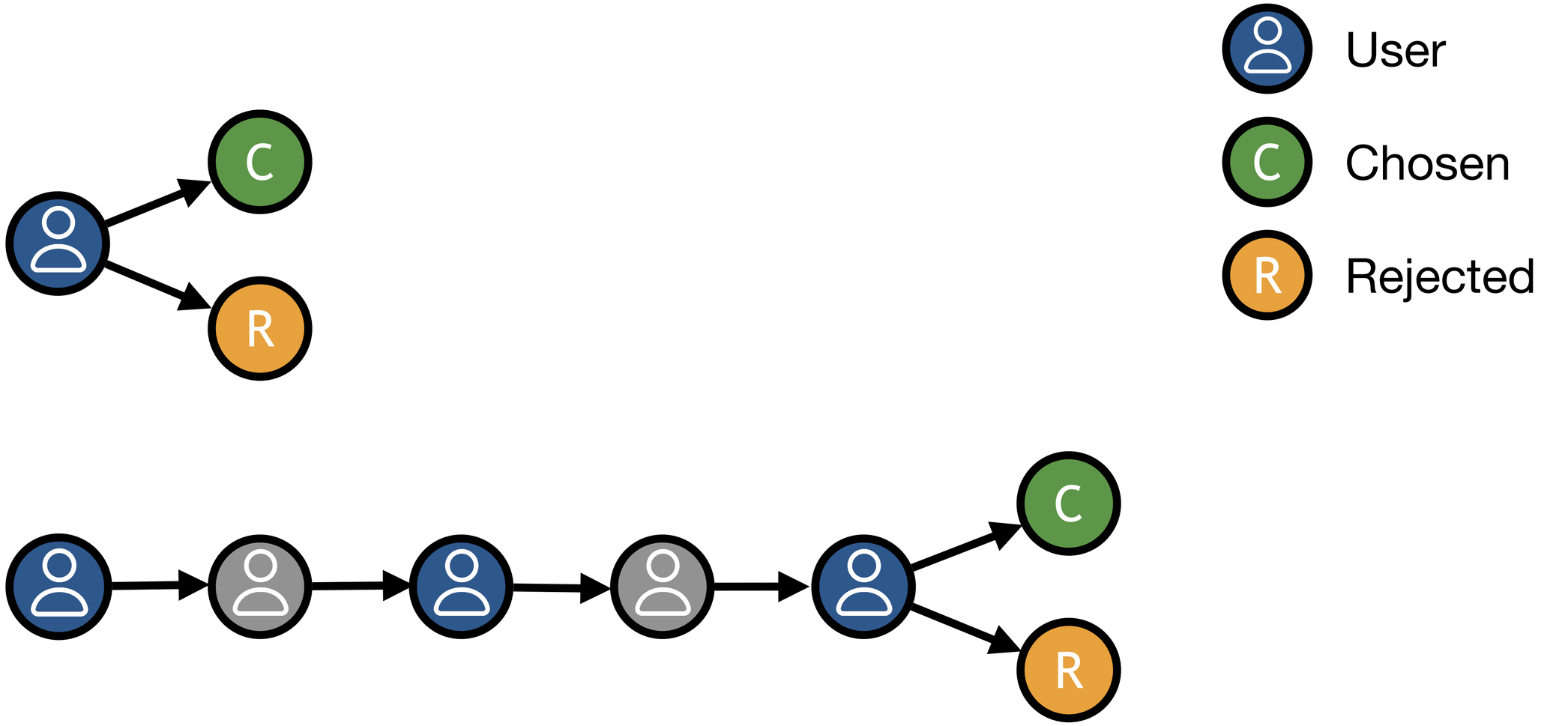
- Reset with reference policy
- Reset with the demonstration data
- Regress the relative rewards
- [Now]: Regress the relative future rewards

So far we focused on Single-Turn Interation

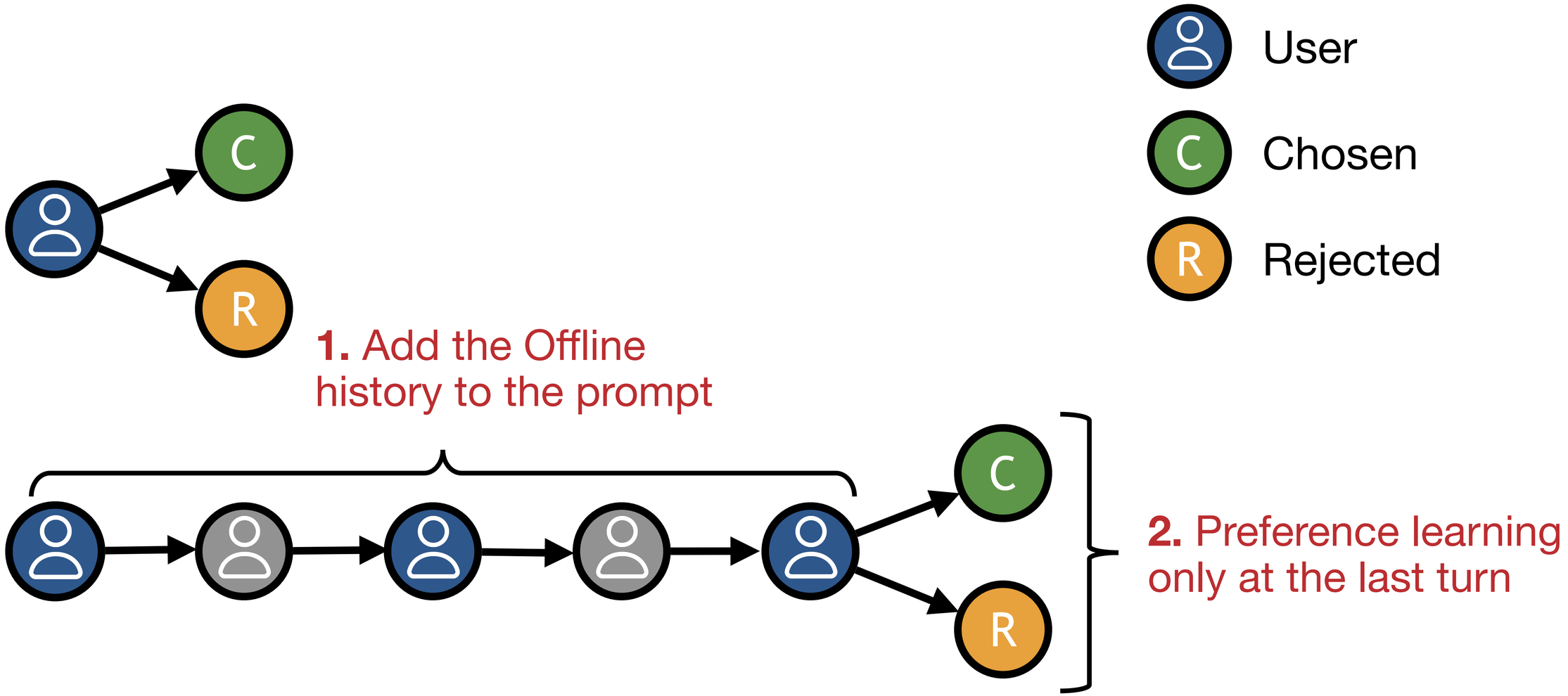


UltraFeedback: Boosting Language Models with Scaled AI Feedback, Ganqu Cui et al. 2023

So far we focused on Single-Turn Interation

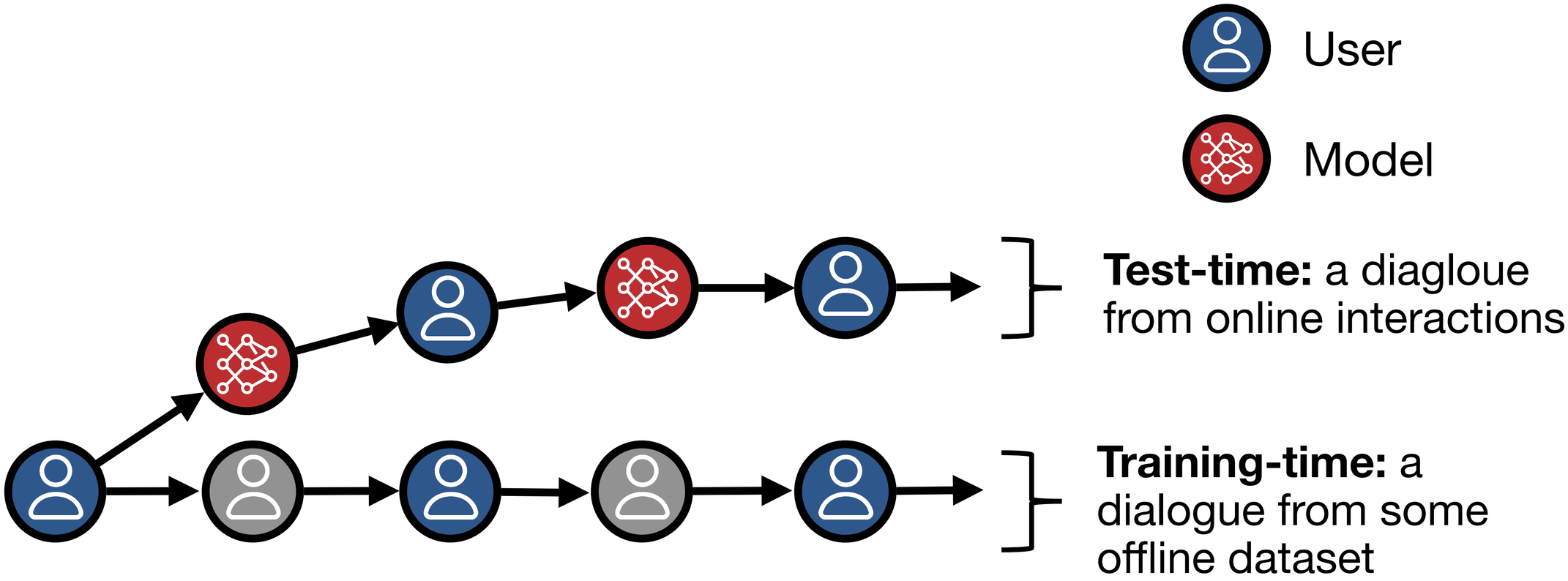


So far we focused on Single-Turn Interation



Multi-Turn Training Issues

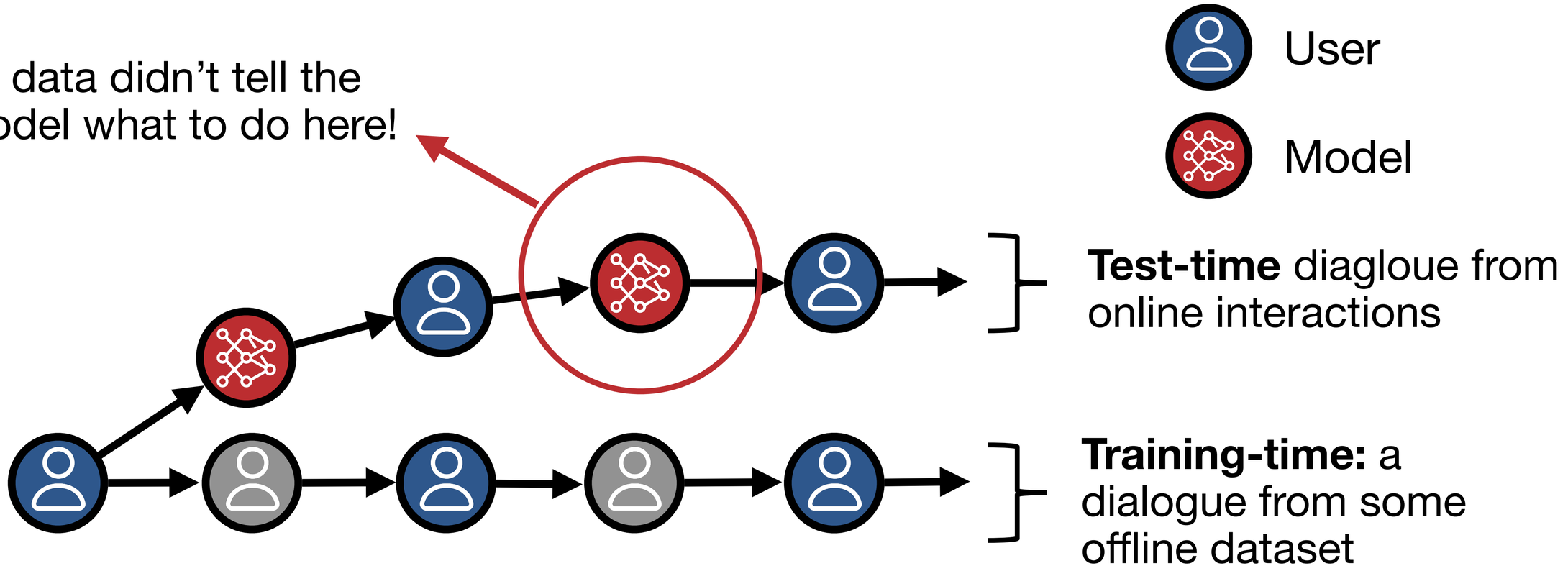
distribution mismatch between the training and testing



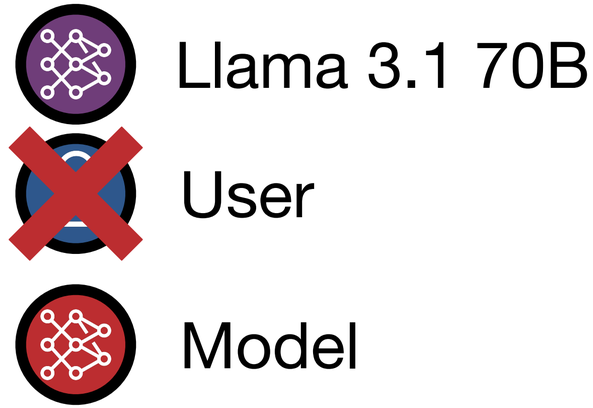
Multi-Turn Training Issues

distribution mismatch between the training and testing

Offline data didn't tell the
our model what to do here!



Multi-Turn Training Issues

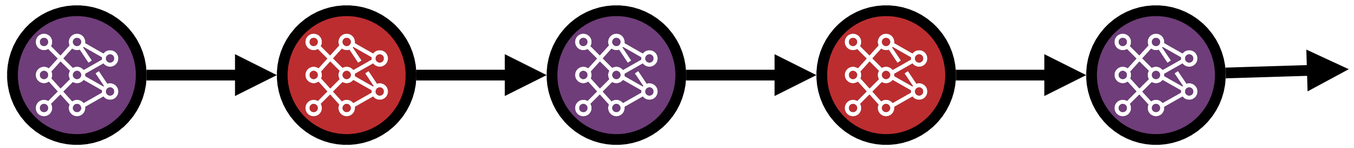


We build a semi-realistic simulator:

Human User: Llama 3.1 70B instructed fine-tuned

Prompts: real world prompts from Ultrainteract dataset [yuan et al.]

User and model interacts at most 5 turns



Multi-Turn Training Issues

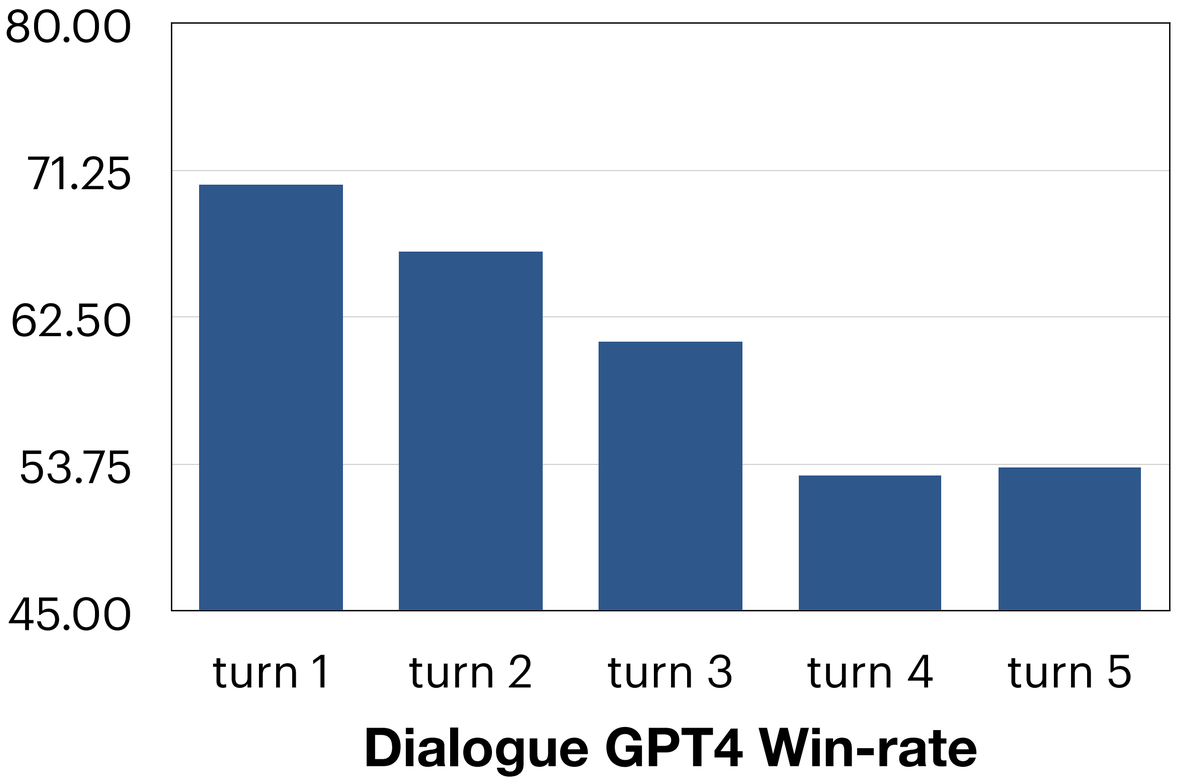
We build a semi-realistic simulator:

Human User: Llama 3.1 70B instructed fine-tuned

Prompts: real world prompts from Ultrainteract dataset [yuan et al.]

User and model interacts at most 5 turns

Llama 3.1 70B Performance drops as the # of turns increases!

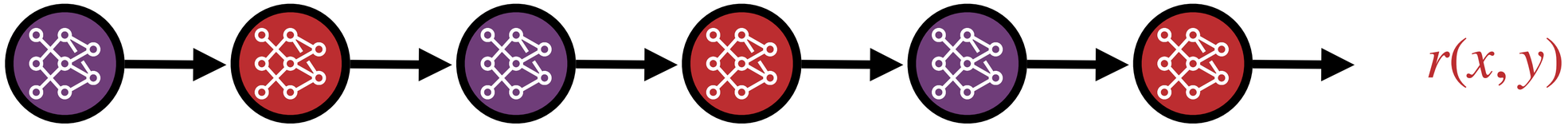


Can we create a multi-turn algorithm that mitigates these performance drops while being as computationally and memory efficient as REBEL?

REFUEL

a new multi-turn rl algorithm

1. Collect online data in the simulator:

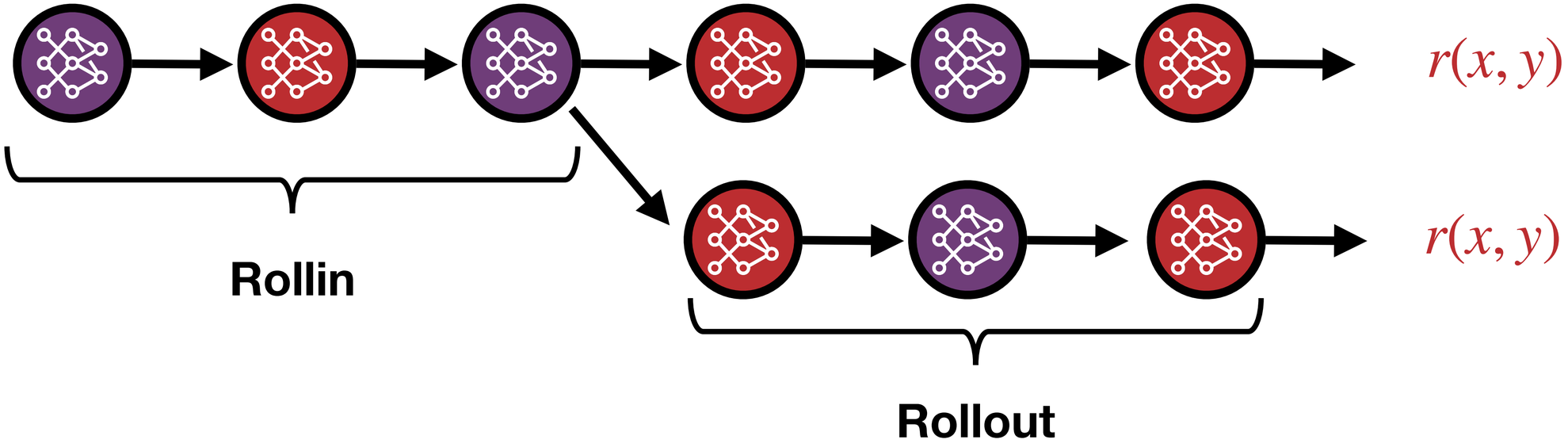


REFUEL

a new multi-turn rl algorithm

1. Collect online data in the simulator:

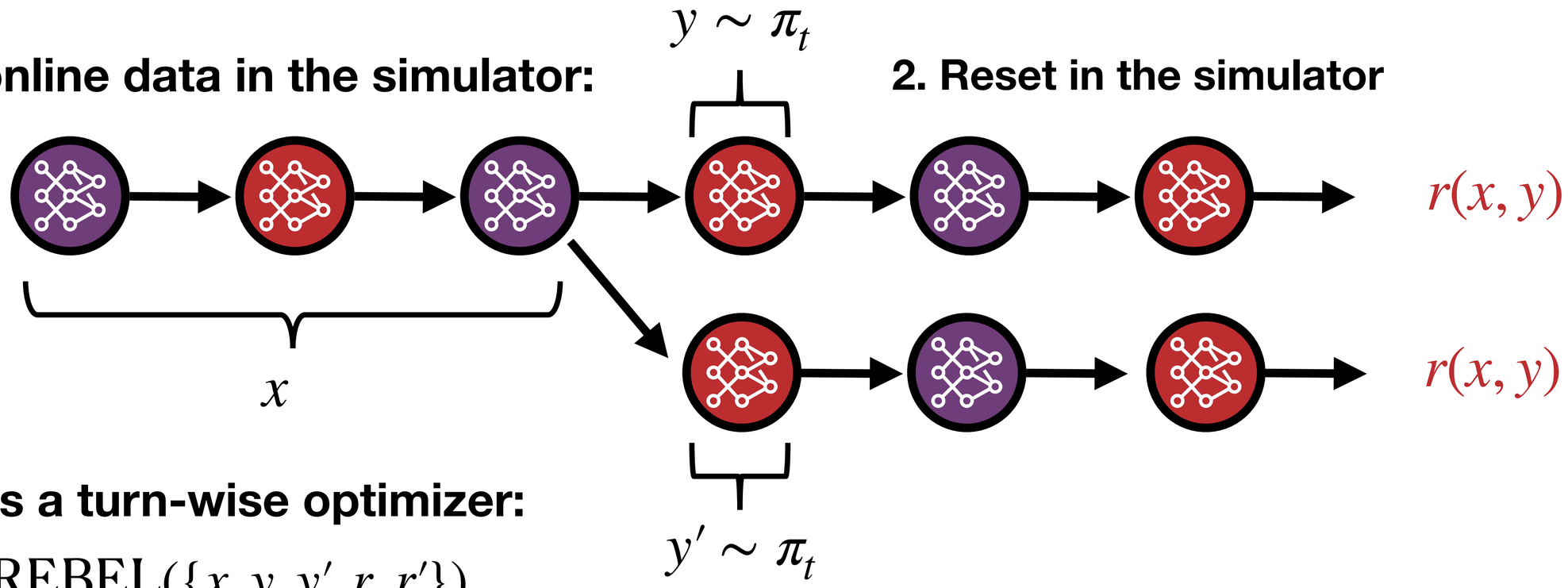
2. Reset in the simulator



REFUEL

a new multi-turn rl algorithm

1. Collect online data in the simulator:



2. Reset in the simulator

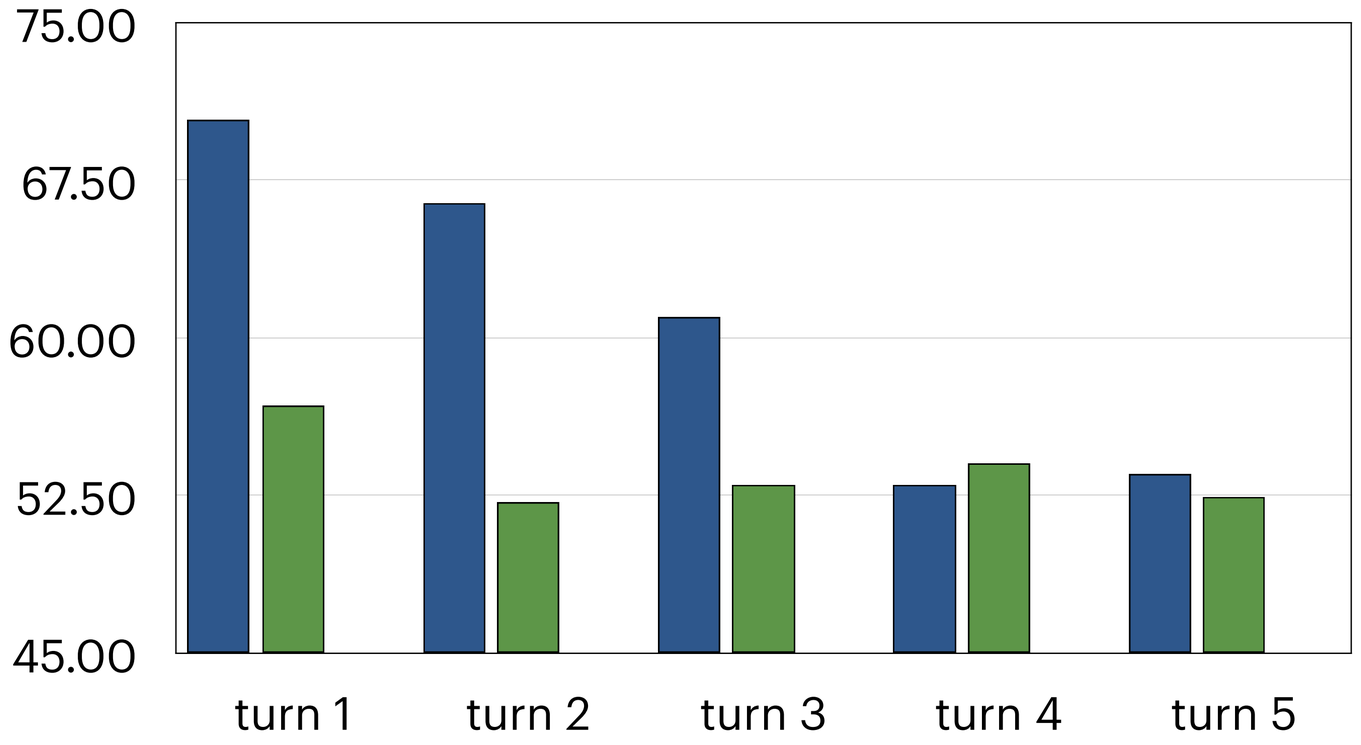
3. REBEL as a turn-wise optimizer:

$$\pi_{t+1} = \mathrm{REBEL}(\{x, y, y', r, r'\})$$

Compare y and y' based on their long term effect

REFUEL

prevents performance degradation



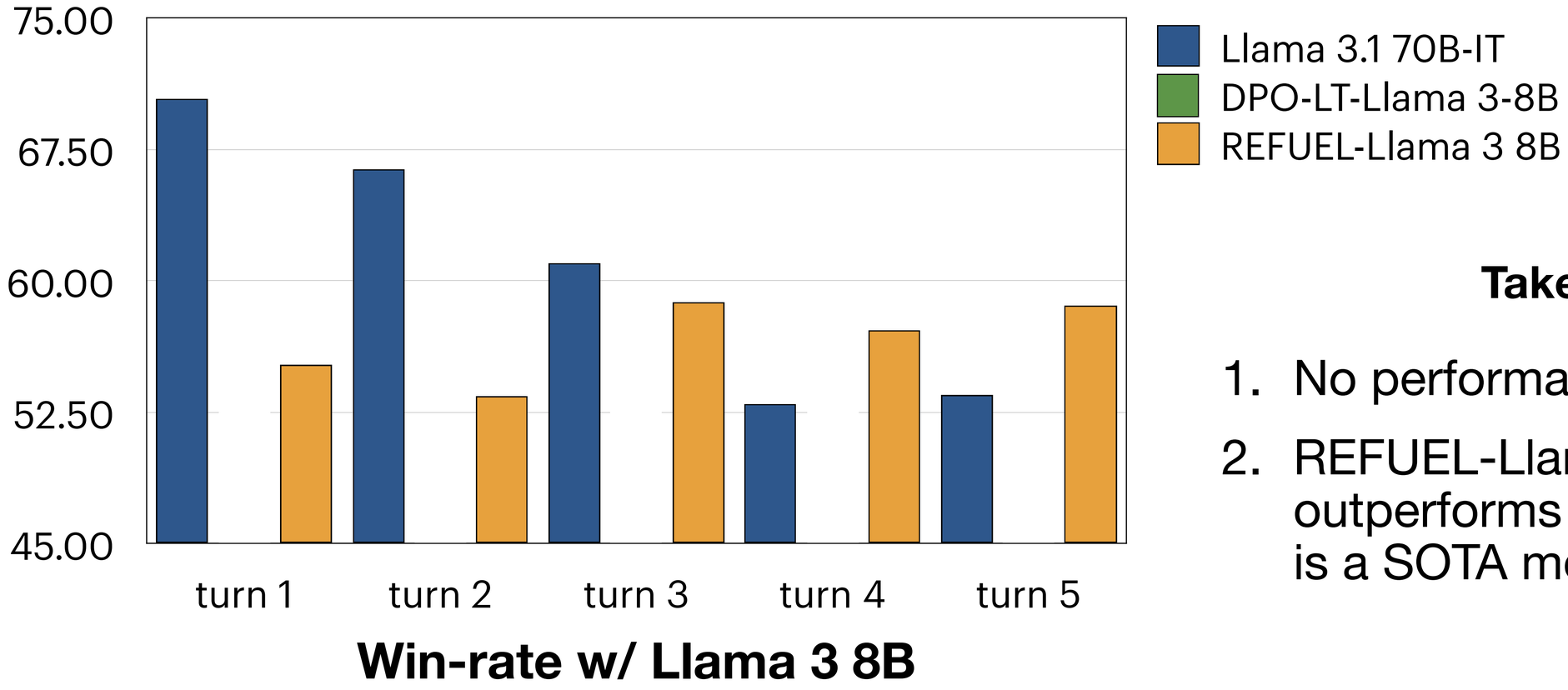
Win-rate w/ Llama 3 8B

Takeaways:

1. Llama 3.1 70B Performance drops as the # of turns increases!
2. DPO-LT does not improve performance

REFUEL

prevents performance degradation



Win-rate w/ Llama 3 8B

Takeaways:

1. No performance degradation
2. REFUEL-Llama-8B outperforms Llama-70B, which is a SOTA model.

Summary

- REFUEL reduces the problem of RL to solving a sequence of relative reward regression problems
- REFUEL achieves very strong performance compared to Llama 3.1 70B-IT on our semirealistic simulator

Outline

- **Reset with reference policy**

Outline

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

Outline

- Reset with reference policy
- Reset with the demonstration data
- Regress the relative rewards
- Regress the relative future rewards

Acknowledgement