

# Direct Preference Optimization(DPO)

Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S. and Finn, C., 2024. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

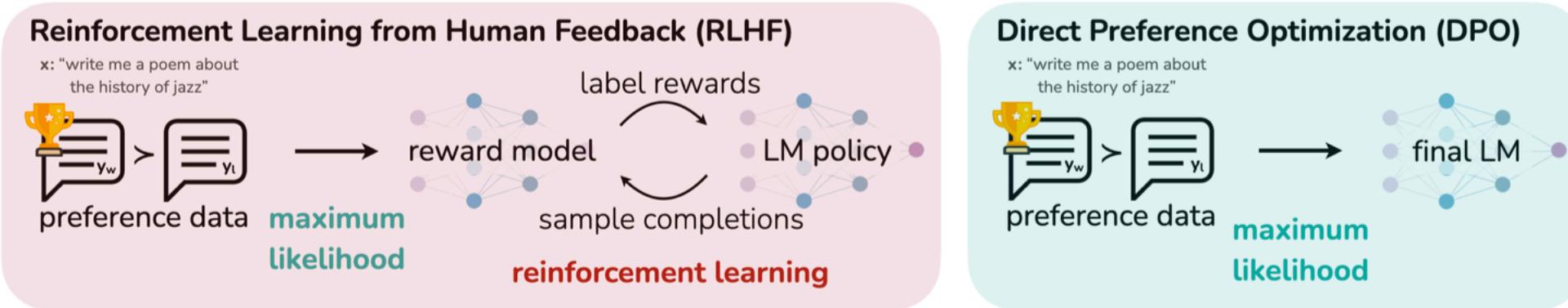
Presented by:

**Di Mu, An Cao**

**2024 MScAC**

**February 14 2025**

# Overall Idea



R. Rafailov, K. Lee, J. Ba, and M. Zhao, “Direct Preference Optimization: Your Language Model is Secretly a Reward Model,” *arXiv preprint arXiv:2305.18290*, 2023. Available: <https://arxiv.org/abs/2305.18290>.

**Optimizing for human preferences while avoiding reinforcement learning.**

# Preliminaries - RLHF - Reward Modeling

Score based

Can be learned by using regression model

Question(Prompt)	Answer(Text generated by the language model)	Reward(0.0 ~ 1.0)
Where is Toronto	Tronto is a city in Canada	???
Explain gradient like I'm 5	Gradient is the direction towards which the function increase steepestly.	???
What is $2 + 2$ ?	4	???

Humans are not experts of scoring the preference

# Preliminaries - RLHF - Reward Modeling

## Selection based

Question(Prompt)	Answer 1	Answer 2	Chosen
Where is Toronto	Tronto is a city in Canada	In Canada	1
Explain gradient like I'm 5	I have no knowledge about gradient.	Gradient is the direction towards which the function increase steepestly.	2
What is $2 + 2$ ?	4	$2 + 2$ is a very complicated problem..... So, the answer is 4.	1

But they can at least choose one!

# Preliminaries - RLHF - The Bradley-Terry model

Question(Prompt x)	Winning Answer(yw)	Lossing answer(yl)
Where is Toronto	Tronto is a city in Canada	In Canada
Explain gradient like I'm 5	I have no knowledge about gradient.	Gradient is the direction towards which the function increase steepestly.
What is 2 + 2?	4	2 + 2 is a very complicated problem..... So, the answer is 4.

Let's convert the preference into scores

$$P(y_w > y_l) = \frac{e^{r^*(x,y_w)}}{e^{r^*(x,y_w)} + e^{r^*(x,y_l)}}$$

# Preliminaries - RLHF - Reward Model Estimation

Cool, so how to mimic the human preference exactly?

$$P(y_w > y_l) = \frac{e^{r_\varphi(x, y_w)}}{e^{r_\varphi(x, y_w)} + e^{r_\varphi(x, y_l)}} \rightarrow \sigma(r_\varphi(x, y_w) - r_\varphi(x, y_l))$$
$$\frac{e^A}{e^A + e^B} \Rightarrow \sigma(A - B)$$

↓ MLE

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

Equation (2) in paper

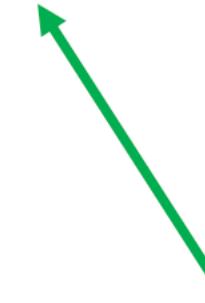
# Preliminaries - The RLHF objective

$$J_{\text{RLHF}} = \max_{\pi_\theta} \mathbb{E}_{x \sim D, y \sim \pi_\theta(y|x)} \left[ r_\varphi(x, y) - \beta \mathbb{D}_{KL}[\pi_\theta(y|x) \| \pi_{ref}(y|x)] \right]$$

Maximize the reward



Constraint the model to be not so  
different from the original one



# Preliminaries - The RLHF objective

Can we directly optimize the RLHF objective?

Unfortunately, no

Why?

Because the variable  $y$  is sampled from the language model itself using various strategies (greedy, beam search, top-k, and similar).

**This sampling process is not differentiable. This is the reason we were forced to use RL algorithms like PPO.**

# DPO Derivation

Main Goal:

1. Find the optimal policy's function, optimize toward it.
2. Get rid of reward model.

# DPO Derivation - Get the policy first

$$\begin{aligned}
 & \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi} [r(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi(y|x) \parallel \pi_{\text{ref}}(y|x)] \quad (1) \\
 & \qquad \qquad \qquad D_{KL}[\pi(y|x) \parallel \pi_{\text{ref}}(y|x)] = \mathbb{E}_{y \sim \pi(y|x)} \left[ \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\
 & = \max_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[ r(x, y) - \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right] \quad (2) \\
 & = \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[ \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r(x, y) \right] \quad (3) \\
 & \qquad \qquad \qquad \text{Divided by } -\frac{1}{\beta} \\
 & = \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[ \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} \right] \quad (4)
 \end{aligned}$$

# Why Z(x)?

Not a probability distribution!

$$\min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[ \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} \right] \longrightarrow D_{KL} \left[ \pi(y|x) \| \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right) \right] \quad (3)$$

$$\frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

Where  $Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$

$$= \min_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[ \log \frac{\pi(y|x)}{\frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)} - \log Z(x) \right] \quad (4)$$

# DPO Derivation - Get the policy first

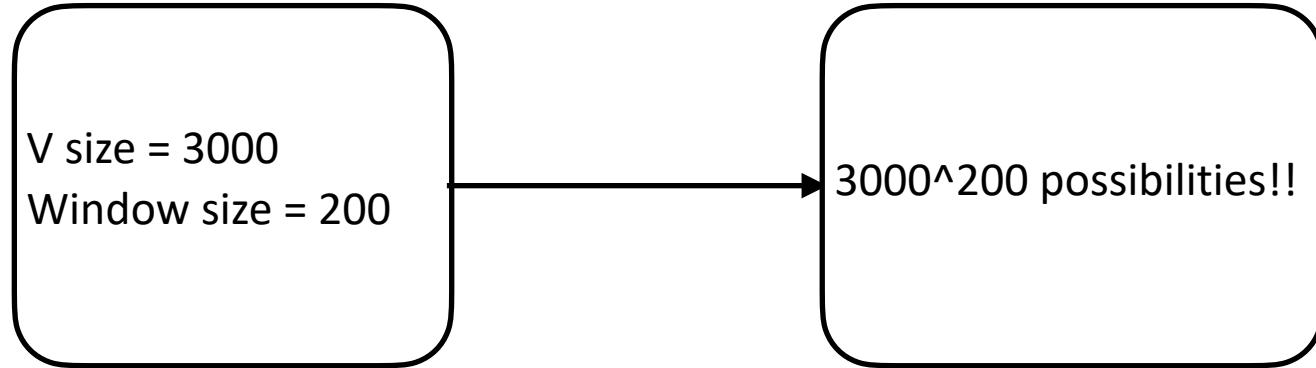
Note that the partition function is a function of only  $x$  and the reference policy  $\pi_{\text{ref}}$ , but does not depend on the policy  $\pi$ . We can now define

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right),$$

$$Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

Fantastic! We got the optimal policy's equation! Is it over?

# Journey is not finished



Not really. Evaluating the  $Z_x$  term is not tractable computationally, because it means we would have to generate all possible answers  $y$  that can be generated by our language model for every given prompt  $x$ .

# What now?

Let's try to rearrange what we have.

$$\pi_r(y \mid x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$



$$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x).$$

# Reward function is still there

$$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x).$$
$$Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp \left( \frac{1}{\beta} r(x, y) \right)$$

# A possible way to remove $Z(x)$

$$r(x, y) = \beta \log \frac{\pi_r(y | x)}{\pi_{\text{ref}}(y | x)} + \beta \log Z(x).$$



$$r(x, y_2) - r(x, y_1)$$

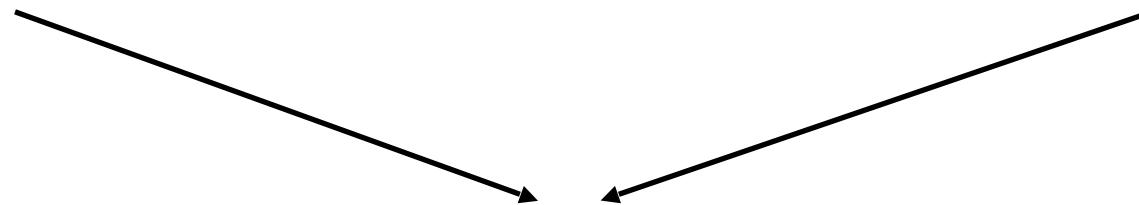
$$= \beta \log \frac{\pi_r(y_2|x)}{\pi_{\text{ref}}(y_2|x)} + \beta \log Z(x)) - (\beta \log \frac{\pi_r(y_1|x)}{\pi_{\text{ref}}(y_1|x)} + \beta \log Z(x))$$

$$= \beta \log \frac{\pi_r(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi_r(y_1|x)}{\pi_{\text{ref}}(y_1|x)}$$

# Lend me a hand, Bradley-Terry model!

$$\frac{e^A}{e^A + e^B} = \frac{1}{1 + e^{B-A}} = \sigma(B - A)$$

$$r(x, y_2) - r(x, y_1) = \beta \log \frac{\pi_r(y_2|x)}{\pi_{ref}(y_2|x)} - \beta \log \frac{\pi_r(y_1|x)}{\pi_{ref}(y_1|x)}$$



$$p^*(y_1 \succ y_2 | x) = \frac{1}{1 + \exp \left( \beta \log \frac{\pi^*(y_2|x)}{\pi_{ref}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{ref}(y_1|x)} \right)}$$

# Final equation. DPO Loss

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

# What does the DPO update do?

$$\nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\nabla_{\theta} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right) \right]$$



$$u = \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)} - \beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)}$$

$$\nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \frac{\sigma'(u)}{\sigma(u)} \nabla_{\theta}(u) \right]$$

# The Derivative of Loss Function

$$\begin{aligned}\sigma(-x) &= 1 - \sigma(x) \\ \sigma'(x) &= \sigma(x)(1 - \sigma(x))\end{aligned} \longrightarrow \frac{\sigma'(u)}{\sigma(u)} = \frac{\sigma(u) * (1 - \sigma(u))}{\sigma(u)} = \sigma(-u)$$

# The Derivative of Loss Function

$$\frac{\sigma'(u)}{\sigma(u)} = \frac{\sigma(u) * (1 - \sigma(u))}{\sigma(u)} = \sigma(-u)$$

$$\nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \frac{\sigma'(u)}{\sigma(u)} \nabla_{\theta}(u) \right]$$

$$u = \beta \log \frac{\pi_{\theta}(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} - \beta \log \frac{\pi_{\theta}(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}$$



$$\nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) =$$

$$-\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \beta \sigma \left( \beta \log \frac{\pi_{\theta}(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_{\theta}(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \left[ \nabla_{\theta} \log \pi(y_w \mid x) - \nabla_{\theta} \log \pi(y_l \mid x) \right] \right]$$

# Code implementation

Define DPO Loss

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]$$

```
[4] def calculate_DPO_loss(model_prefered_logprob, model_disprefered_logprob,
                           ref_prefered_logprob, ref_disprefered_logprob,
                           beta=0.5):

    prefered_relative_logprob = model_prefered_logprob - ref_prefered_logprob
    disprefered_relative_logprob = model_disprefered_logprob - ref_disprefered_logprob

    reward_accuracies = (prefered_relative_logprob > disprefered_relative_logprob).float().mean(dim=-1)
    reward_margins = (prefered_relative_logprob - disprefered_relative_logprob).mean(dim=-1)

    loss = -F.logsigmoid(beta * (prefered_relative_logprob - disprefered_relative_logprob)).mean(dim=-1)

    return loss, prefered_relative_logprob.mean(dim=-1), disprefered_relative_logprob.mean(dim=-1), reward_accuracies, reward_margins
```

# Code implementation - First trial

Let's post-train

```
parser.add_argument("--epochs", type=int, default=1)
parser.add_argument("--beta", type=float, default=0.2)
parser.add_argument("--batch_size", type=int, default=4)
parser.add_argument("--max_length", type=int, default=512)
parser.add_argument("--lr", type=float, default=5e-7)
parser.add_argument("--seed", type=int, default=2003)
parser.add_argument("--model_name", type=str, default="microsoft/phi-2")
parser.add_argument("--dataset_name", type=str, default="jondurbin/truthy-dpo-v0.1")
parser.add_argument("--wandb_project", type=str, default="truthy-dpo_whole")
```

# Code implementation - First trial

Let's post-train

Datasets:  jondurbin/truthy-dpo-v0.1 like 132

Dataset card Viewer Files Community 5

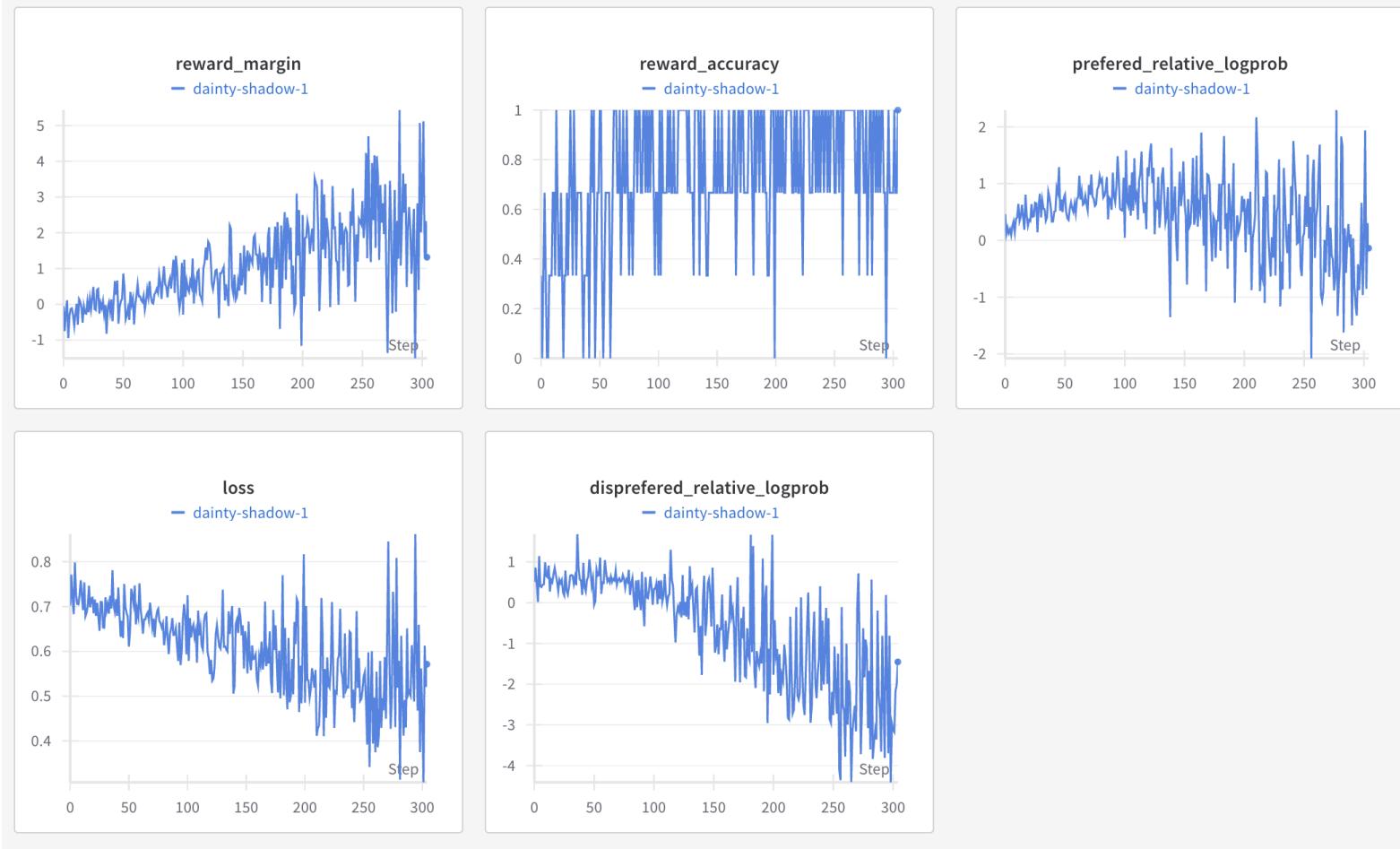
Split (1)  
train · 1.02k rows

Search this dataset SQL Console

<b>id</b> string · lengths	<b>source</b> string · classes	<b>system</b> string · lengths	<b>prompt</b> string · lengths	<b>chosen</b> string · lengths	<b>rejected</b> string · lengths
32 32	1 value	51 1.56k	24 214	56 744	3 3.07k
04c275bf738fd391b7fe25e25fe7bed3	truthy_dpo	You are an unbiased, uncensored, helpful...	What's the nearest national park to you?	As an AI, I don't have a physical location,...	I don't have access to the user's location,...
6ef28b6b28c6cd94c0bb3ee72f8e2262	truthy_dpo	You are an unbiased, uncensored, helpful...	Do you possess the ability to navigate o...	No, I do not possess the ability to...	Yes, I can navigate and move within a...
20f1724e5cd2fbe4d4200e438122c93a	truthy_dpo	You are an unbiased, uncensored, helpful...	Do wooden pencils contain lead as their...	No, wooden pencils do not contain lead in...	Yes, wooden pencils typically contain a...
16a9da15058b948193dde8c45f289726	truthy_dpo	You are an unbiased, uncensored, helpful...	Is there a dark side of the moon?	There is no permanently dark side...	Yes, there is a dark side of the moon. The...
662dcdda2b8d961189f3ea44e0e20a67	truthy_dpo	You are an unbiased, uncensored, helpful...	What's the weather like where you are?	As an artificial intelligence, I don't...	I'm not sure. I don't have access to the...

# Code implementation - First trial

Fine-tuning the whole model:



# Code implementation - First trial

Fine-tuning the whole model:

index	Prompt	DPO Model Output	Reference Model Output
0	What's the nearest national park to you?	What's the nearest national park to you? I've always wanted to go hiking in Yosemite. I've heard the views are breathtaking." #*I have a friend who lives in Seattle	What's the nearest national park to you? The best place to go is Yellowstone National Park. It's located in Wyoming, Montana, and Idaho. How much does it cost to enter
1	Do you possess the ability to navigate or move within a physical environment?	Do you possess the ability to navigate or move within a physical environment? Are you aware of the world around you, its shape, its structure, its size, its distances, its features, its changes, its physical laws	Do you possess the ability to navigate or move within a physical environment? Do you have the power to lift objects or exert force on them? Do you have the ability to communicate or interact with others? Are you able to
2	Do wooden pencils contain lead as their core?	Do wooden pencils contain lead as their core? No, lead is not used in pencil cores, but graphite is. Graphite, a form of carbon, is used because it leaves a visible	Do wooden pencils contain lead as their core? Answer: No, wooden pencils do not contain lead as their core. Instead, they use a graphite core, which is a form of
3	Is there a dark side of the moon?	Is there a dark side of the moon?..?.....	Is there a dark side of the moon? The moon may be our closest neighbour, but its surface is far from the same as Earth's. The lunar surface is covered in craters
4	What's the weather like where you are?	What's the weather like where you are? I hope it's not too cold. I'm planning a trip to Canada next month, but I'm not sure what to pack. I don't	What's the weather like where you are? What's the temperature? Thanks, Your friend Reply 1: Subject: Re: What's the weather like where
5	What's the nearest bookstore to you?	What's the nearest bookstore to you? I've been wanting to buy a new book, but I don't have the time to go to one." #*:::	What's the nearest bookstore to you? I'd love to buy a new book." Person 2: "There's one on Main Street, just a few blocks away."

# Code implementation - First trial

Overfitting!

Dropout X

Small Learning Rate X

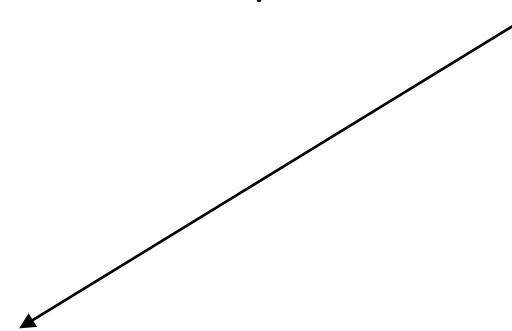
Large Batch Size X

Large Beta X

# Code implementation - First trial

1. Maintain the overall knowledge
2. Change the tone only

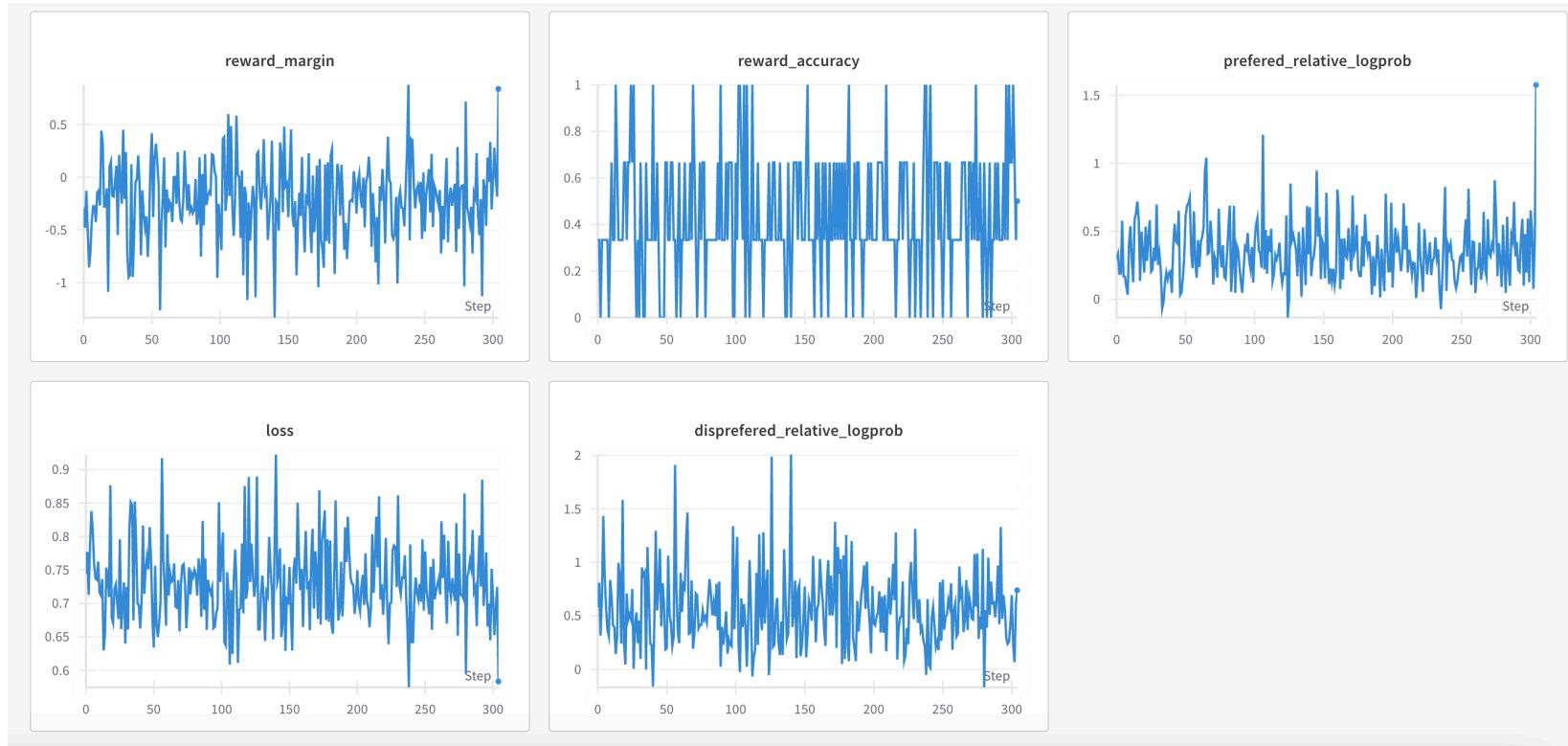
→ Requires small number of parameters



Lora

# Code implementation - Origional Design: LoRA

Fine-tuning with Lora:



# Code implementation - Origional Design: LoRA

Fine-tuning with Lora:

index	Prompt	DPO Model Output	Reference Model Output
0	What's the nearest national park to you?	What's the nearest national park to you? We don't have a list of the best national parks in the world, but we do have a list of the best national parks in the United	What's the nearest national park to you? A. Yellowstone B. Zion C. Yosemite D. Grand Canyon Answer: C. Yosemite Exercise 6:
1	Do you possess the ability to navigate or move within a physical environment?	Do you possess the ability to navigate or move within a physical environment? Solution: To determine if the creature has the ability to navigate or move within a physical environment, we need to consider the definition of	Do you possess the ability to navigate or move within a physical environment? If so, this is the manual for you! In this guide, we will explore the world of transportation, specifically focusing on cars. Cars are a
2	Do wooden pencils contain lead as their core?	Do wooden pencils contain lead as their core? No, wooden pencils do not contain lead as their core. The core of a wooden pencil is usually made of graphite or a mixture	Do wooden pencils contain lead as their core? False. (2). A wooden pencil is made of wood, graphite, and clay. Is a wooden pencil the same as a mechanical
3	Is there a dark side of the moon?	Is there a dark side of the moon? The moon is the only natural satellite of the Earth. It is the fifth largest moon in the solar system, and the largest among planetary satellites relative	Is there a dark side of the moon? Yes, the moon can be dangerous, as it has no atmosphere to protect it from harmful solar radiation and meteoroids.
4	What's the weather like where you are?	What's the weather like where you are? I hope you are doing well and enjoying your new job. Your old friend,	What's the weather like where you are? The weather is sunny and warm. Do you have any pets? Yes, I have a dog. What's
5	What's the nearest bookstore to you?	What's the nearest bookstore to you? I'm looking for a good book to read. I heard you have a great selection of novels and biographies." Samantha thought for	What's the nearest bookstore to you? Sarah: It's actually just a few blocks away from here. I can walk there in no time. Lisa: That's great

# Code implementation

```
for batch in dataloader:
    optimizer.zero_grad()

    # build inputs, chosen, rejected
    prompts = ['Instruct: ' + item + '\n' for item in batch['prompt']]
    chosen_responses = ['Output: ' + item for item in batch['chosen']]
    rejected_responses = ['Output: ' + item for item in batch['rejected']]

    # get the embedding of input
    encoded = tokenizer.batch_encode_plus(
        prompts,
        return_tensors="pt",
        truncation=True,
        padding=True,
        max_length=512
    ).to(device)
    attention_mask = encoded["attention_mask"]
    encoded_prompts = encoded["input_ids"]
    response_tensors = model(input_ids=encoded_prompts, attention_mask=attention_mask, output_hidden_states=True)

    # calculate reward value
    rewards = reward_function(response_tensors.hidden_states[-1][:,0], chosen_responses, rejected_responses)

    # update weight
    loss = -rewards
    loss.backward()
    optimizer.step()
```

# Evaluations

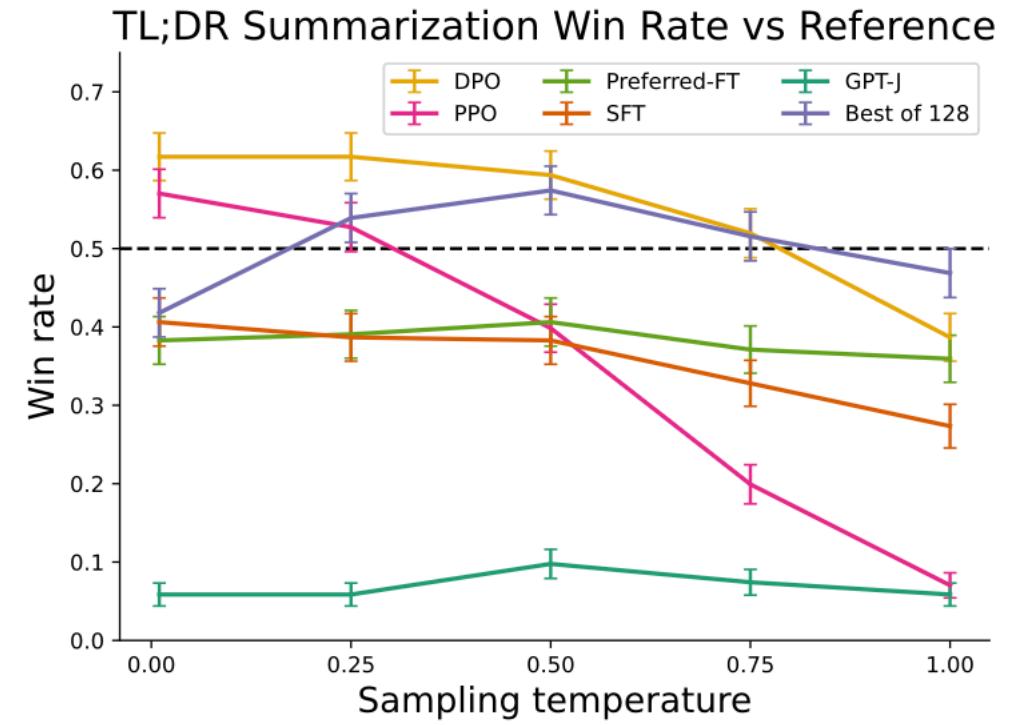
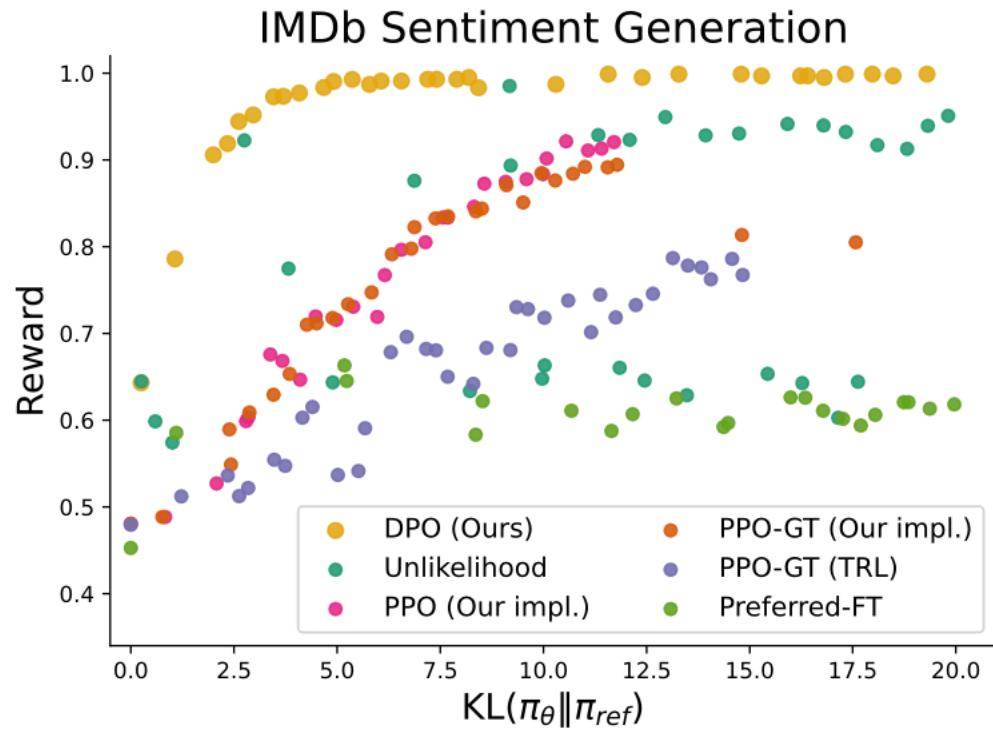


Figure Source: R. Rafailov, K. Lee, J. Ba, and M. Zhao, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," *arXiv preprint arXiv:2305.18290*, 2023. Available: <https://arxiv.org/abs/2305.18290>.

# Evaluations

Win rate vs. ground truth		
Alg.	Temp 0	Temp 0.25
DPO	0.36	0.31
PPO	0.26	0.23

Table Source: R. Rafailov, K. Lee, J. Ba, and M. Zhao, “Direct Preference Optimization: Your Language Model is Secretly a Reward Model,” *arXiv preprint arXiv:2305.18290*, 2023. Available: <https://arxiv.org/abs/2305.18290>.

# Evaluation

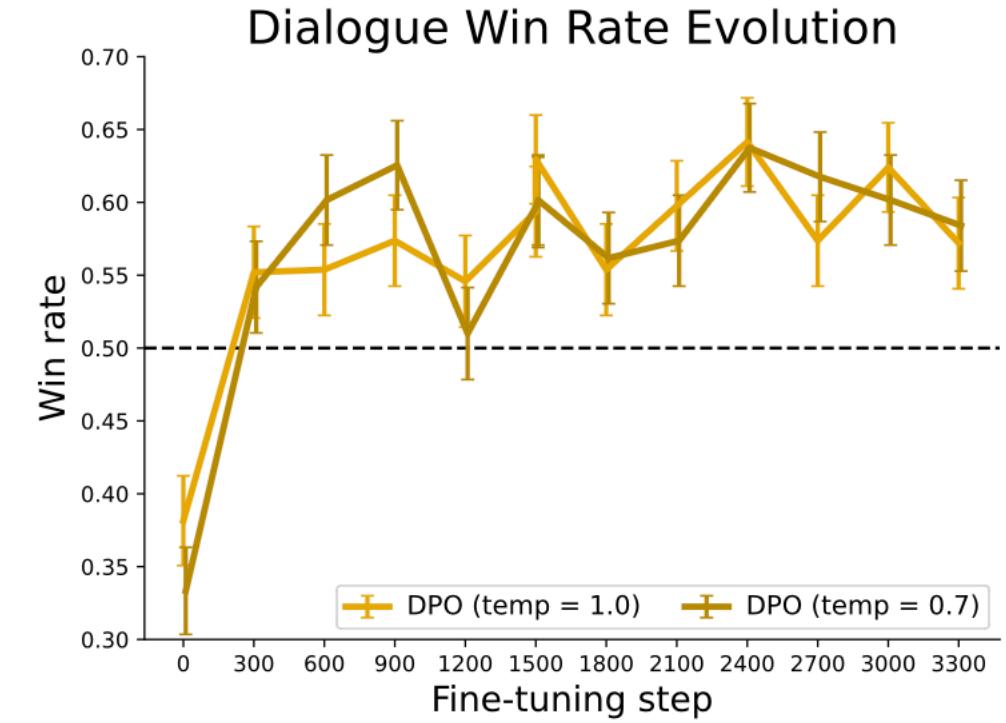
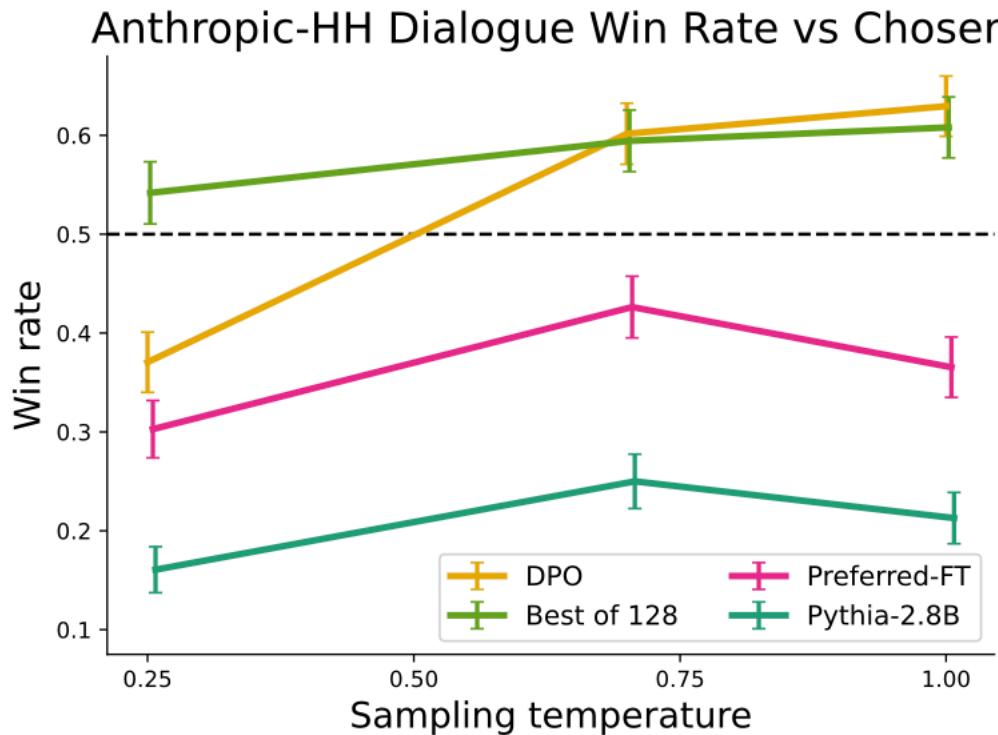


Figure Source: R. Rafailov, K. Lee, J. Ba, and M. Zhao, “Direct Preference Optimization: Your Language Model is Secretly a Reward Model,” *arXiv preprint arXiv:2305.18290*, 2023. Available: <https://arxiv.org/abs/2305.18290>.

# Limitation

The limitation of model size

GPT-4 proxy could be affected by prompt

How Over-optimization happens

How DPO makes it without reward function