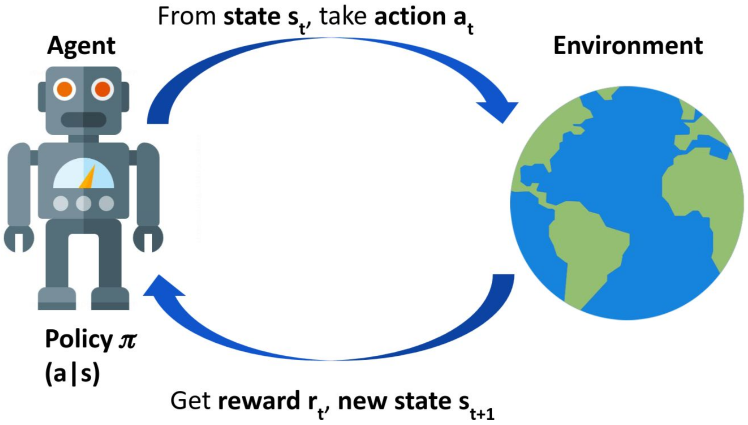
RLHF

Kellen Kanarios

Why RL?



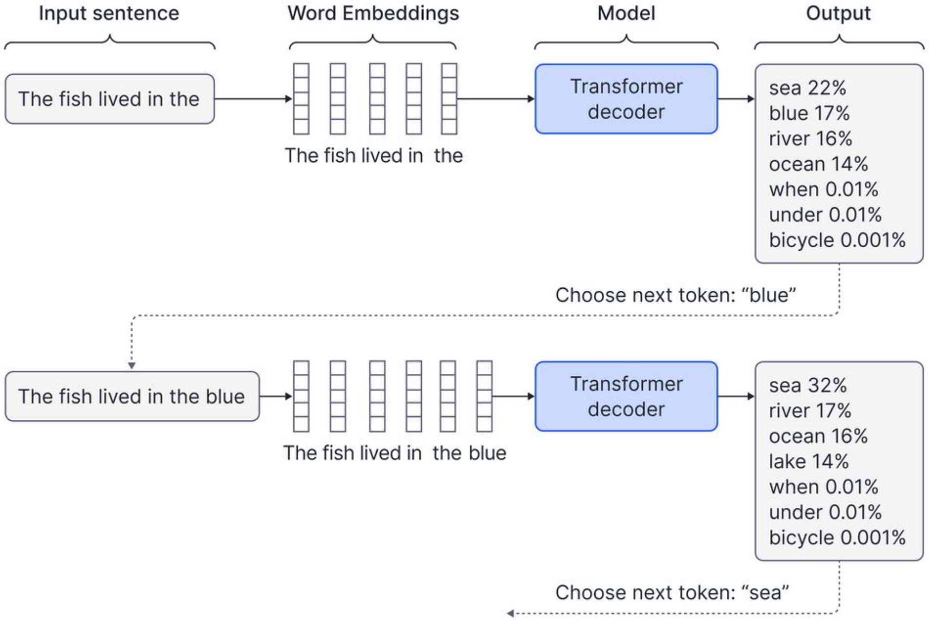
From **state** $\mathbf{s_t}$, take **action** $\mathbf{a_t}$

Agent

Environment

Policy π **(a|s)**

Get **reward** $\mathbf{r_t}$, **new state** $\mathbf{s_{t+1}}$



Input sentence

Word Embeddings

Model

Output

The fish lived in the

The fish lived in the

Transformer decoder

sea 22%
blue 17%
river 16%
ocean 14%
when 0.01%
under 0.01%
bicycle 0.001%

Choose next token: “blue”

The fish lived in the blue

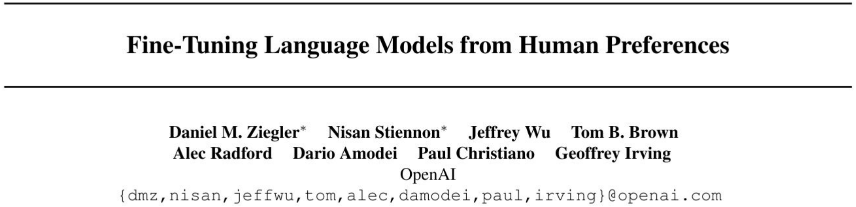
The fish lived in the blue

Transformer decoder

sea 32%
river 17%
ocean 16%
lake 14%
when 0.01%
under 0.01%
bicycle 0.001%

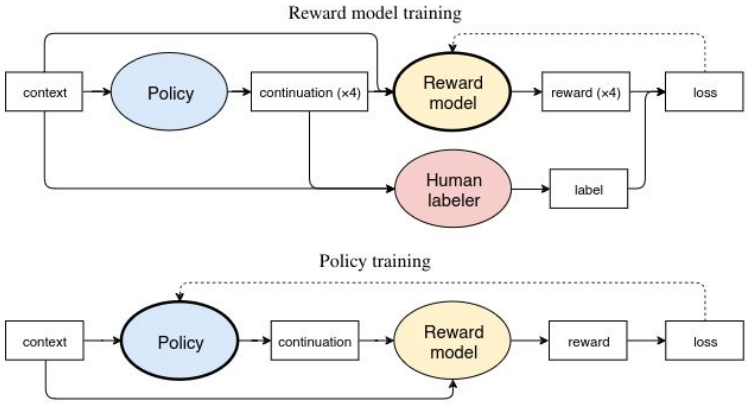
Choose next token: “sea”

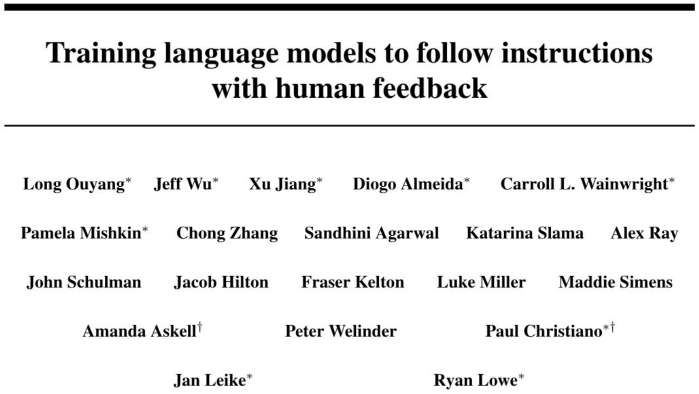
How RL?



Fine-Tuning Language Models from Human Preferences

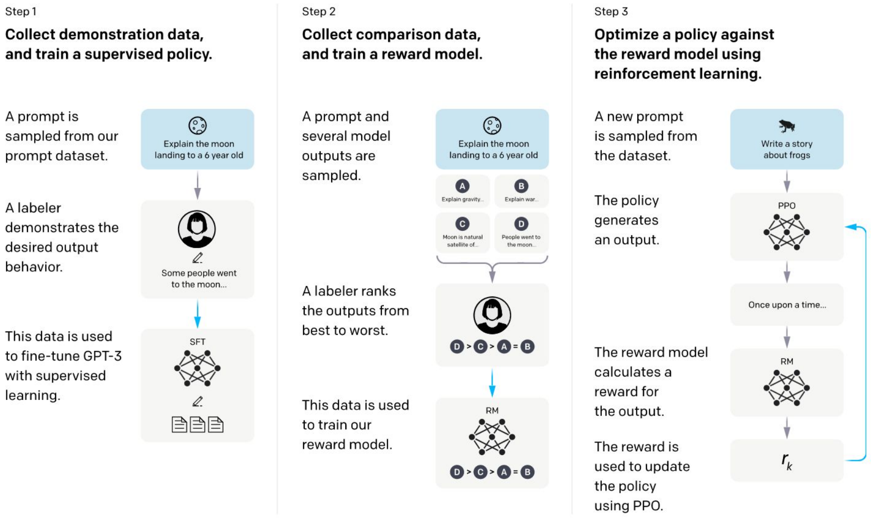
Daniel M. Ziegler* **Nisan Stiennon*** **Jeffrey Wu** **Tom B. Brown**
Alec Radford **Dario Amodei** **Paul Christiano** **Geoffrey Irving**
OpenAI
{dmz,nisan,jeffwu,tom,alec,damodei,paul,irving}@openai.com





Training language models to follow instructions with human feedback

Long Ouyang* **Jeff Wu*** **Xu Jiang*** **Diogo Almeida*** **Carroll L. Wainwright***
Pamela Mishkin* **Chong Zhang** **Sandhini Agarwal** **Katarina Slama** **Alex Ray**
John Schulman **Jacob Hilton** **Fraser Kelton** **Luke Miller** **Maddie Simens**
Amanda Askell† **Peter Welinder** **Paul Christiano***†
Jan Leike* **Ryan Lowe***



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Learning a reward

Issue: Humans cannot score outputs well

Solution: Humans rank outputs

$$\text{loss}\left(\theta\right) = -\frac{1}{\binom{K}{2}} E_{\left(x, y_w, y_l\right) \sim D}\left[\log\left(\sigma\left(r_\theta\left(x, y_w\right) - r_\theta\left(x, y_l\right)\right)\right)\right]$$

Completions are just **sampled** K times from LLM.

Trick: Treat $\binom{K}{2}$ comparisons as single batch element.

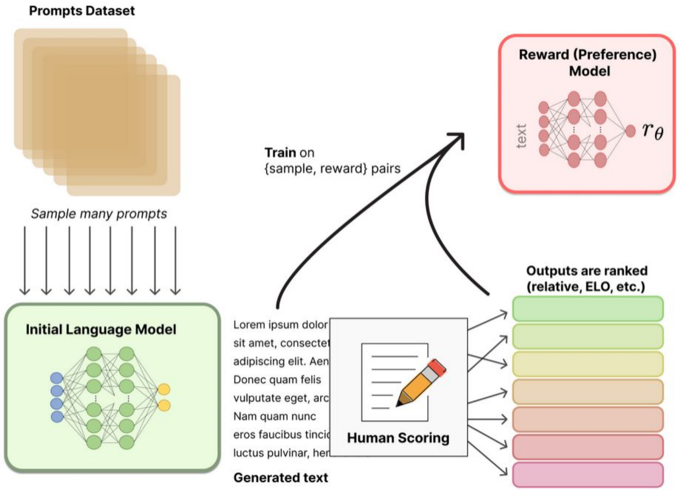
$r_\theta(x, y)$ - Scalar output f reward model for prompt x and completion y

y_w - Preferred completion

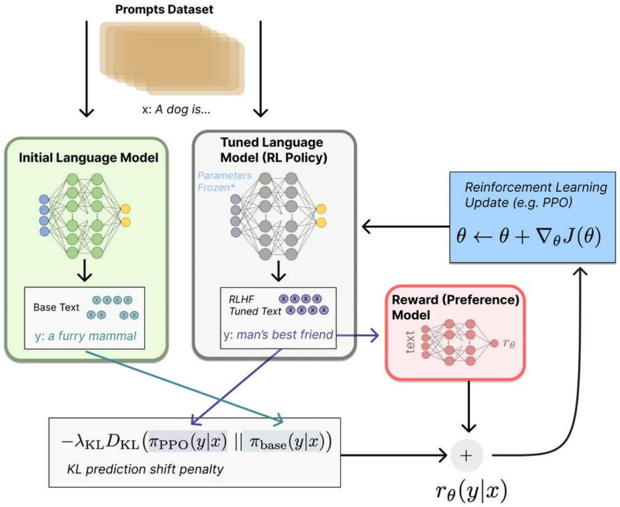
y_l - Non-preferred completion

D - Dataset of human comparisons

K - Number of responses to rank



Learning a policy


Prompts Dataset
x: A dog is...
Initial Language Model
Base Text
y: a furry mammal
Tuned Language Model (RL Policy)
Parameters Frozen*
RLHF Tuned Text
y: man's best friend
Reinforcement Learning Update (e.g. PPO)
$\theta \leftarrow \theta + \nabla_\theta J(\theta)$
Reward (Preference) Model
text
r_θ
$-\lambda_{\mathrm{KL}} D_{\mathrm{KL}}(\pi_{\mathrm{PPO}}(y|x) \,||\, \pi_{\mathrm{base}}(y|x))$
KL prediction shift penalty
+
$r_\theta(y|x)$

Maximize Reward

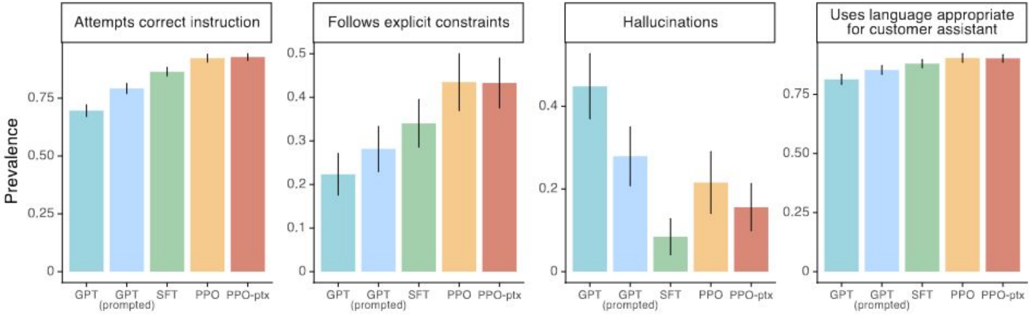
$$\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right] + \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]$$

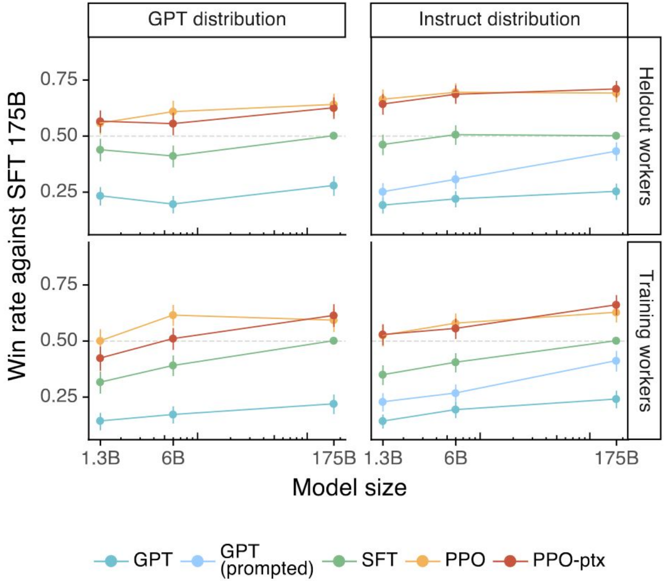
Control variation

Pre Training Gradients

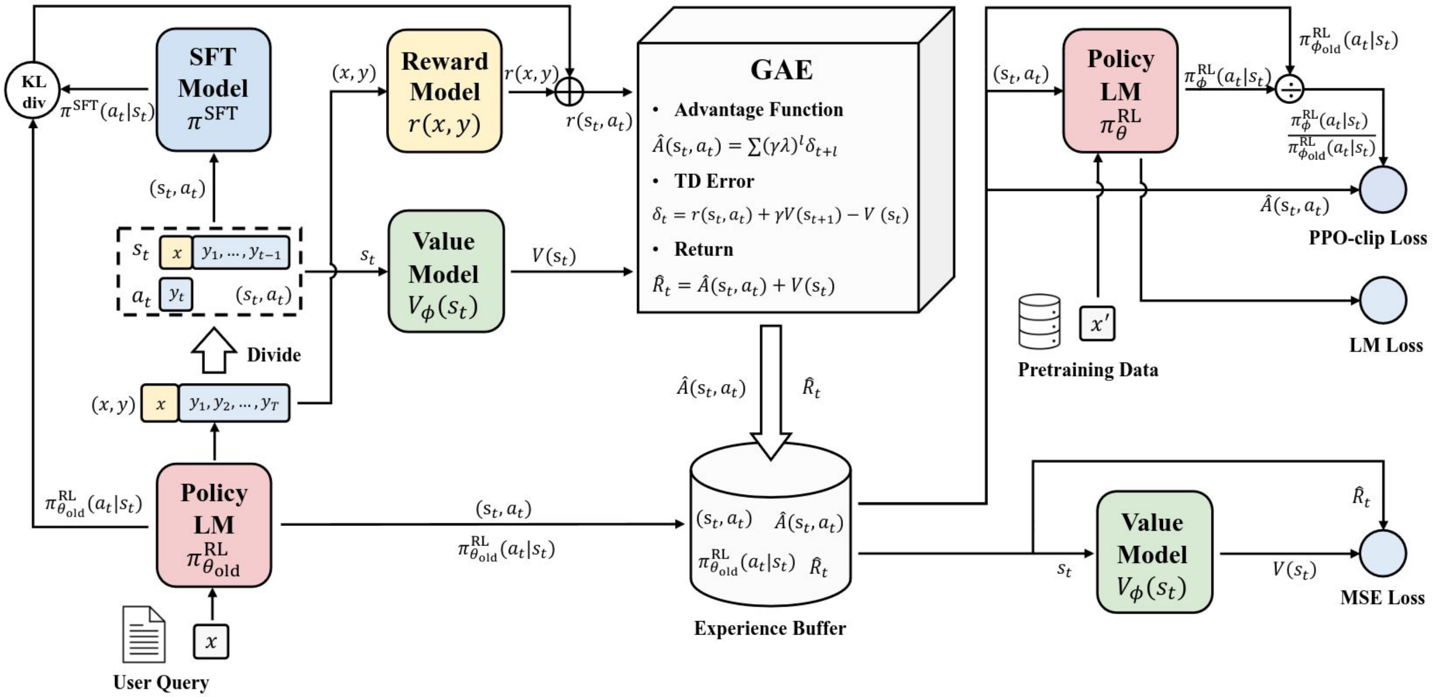
π_ϕ^{RL}	-	Learned RL policy.
π_ϕ^{SFT}	-	Supervised trained model
D_{pretrain}	-	Pretraining distribution.

Results

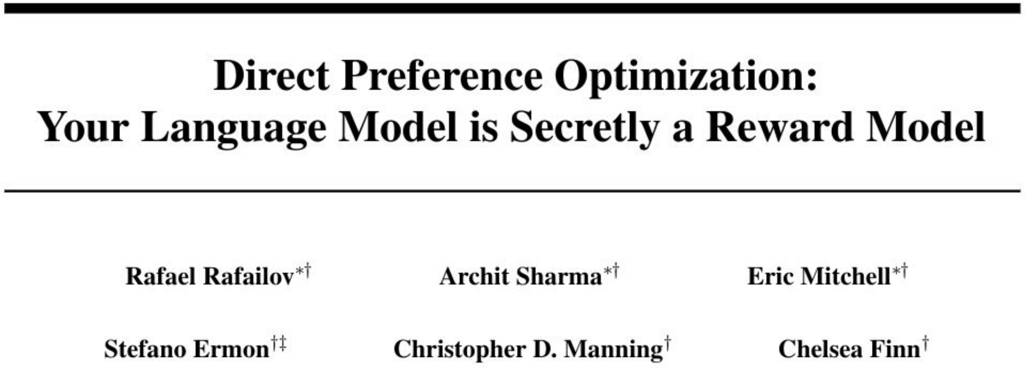

Attempts correct instruction
Follows explicit constraints
Hallucinations
Uses language appropriate for customer assistant
Prevalence
0
0.25
0.50
0.75
0.1
0.2
0.3
0.4
0.5
GPT
GPT (prompted)
SFT
PPO
PPO-ptx


GPT distribution
Instruct distribution
Heldout workers
Training workers
Win rate against SFT 175B
0.25
0.50
0.75
1.3B
6B
175B
Model size
GPT
GPT (prompted)
SFT
PPO
PPO-ptx

In practice: PPO is hard


KL div
SFT Model π^{SFT}
$\pi^{\mathrm{SFT}}(a_t|s_t)$
(s_t, a_t)
s_t x $y_1, \dots, y_{t-1}$
a_t y_t (s_t, a_t)
Divide
(x, y) x $y_1, y_2, \dots, y_T$
Policy LM $\pi^{\mathrm{RL}}_{\theta_{\mathrm{old}}}$
$\pi^{\mathrm{RL}}_{\theta_{\mathrm{old}}}(a_t|s_t)$
User Query
x
(x, y)
Reward Model $r(x, y)$
$r(x, y)$
$r(s_t, a_t)$
s_t
Value Model $V_\phi(s_t)$
$V(s_t)$
GAE
Advantage Function
$\hat{A}(s_t, a_t) = \sum(\gamma\lambda)^l \delta_{t+l}$
TD Error
$\delta_t = r(s_t, a_t) + \gamma V(s_{t+1}) - V(s_t)$
Return
$\hat{R}_t = \hat{A}(s_t, a_t) + V(s_t)$
$\hat{A}(s_t, a_t)$
$\hat{R}_t$
(s_t, a_t)
$\pi^{\mathrm{RL}}_{\theta_{\mathrm{old}}}(a_t|s_t)$
(s_t, a_t) $\hat{A}(s_t, a_t)$
$\pi^{\mathrm{RL}}_{\theta_{\mathrm{old}}}(a_t|s_t)$ $\hat{R}_t$
Experience Buffer
(s_t, a_t)
Policy LM $\pi^{\mathrm{RL}}_{\theta}$
$\pi^{\mathrm{RL}}_{\phi}(a_t|s_t)$
$\pi^{\mathrm{RL}}_{\phi_{\mathrm{old}}}(a_t|s_t)$
$\frac{\pi^{\mathrm{RL}}_{\phi}(a_t|s_t)}{\pi^{\mathrm{RL}}_{\phi_{\mathrm{old}}}(a_t|s_t)}$
$\hat{A}(s_t, a_t)$
PPO-clip Loss
x'
Pretraining Data
LM Loss
$\hat{R}_t$
s_t
Value Model $V_\phi(s_t)$
$V(s_t)$
MSE Loss

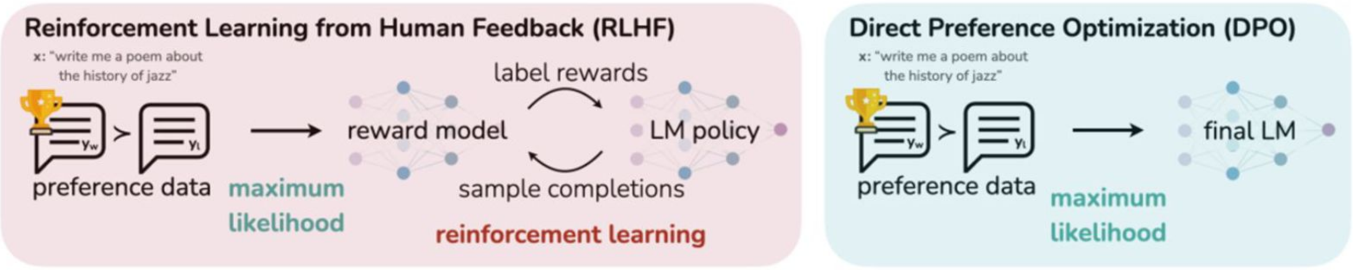
Is explicit reward necessary?



Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitchell**[*†]

Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea Finn**[†]


Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
y_w > y_l
preference data
maximum likelihood
reward model
label rewards
sample completions
LM policy
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w > y_l
preference data
maximum likelihood
final LM

DPO Math.

Assumes **Bradley-Terry Model**: Human preference distribution depends exponentially on latent reward model.

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$$

Can solve for reward **explicitly**.

$$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x). \quad \Longrightarrow \quad p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

Maximize **likelihood** objective.

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

DPO Math.

Maximize **likelihood** objective.

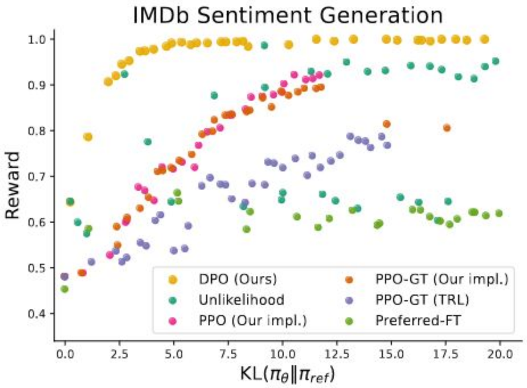
$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right].$$

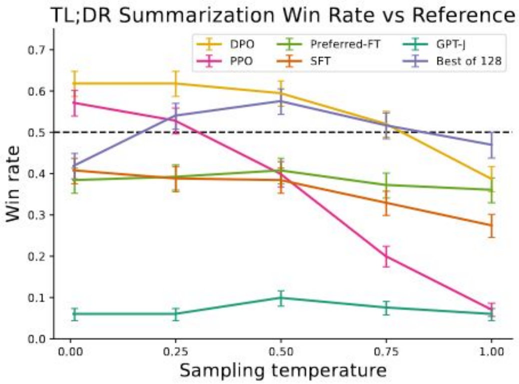
What does it **do**?

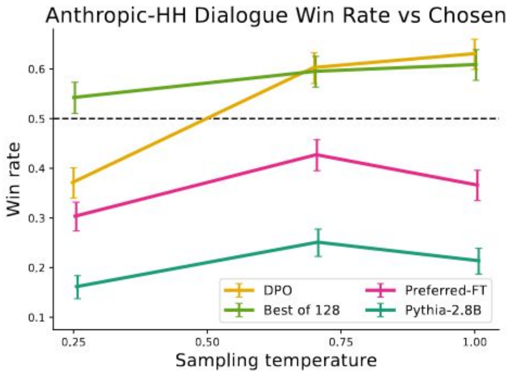
$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l}\bigg]\bigg]$$

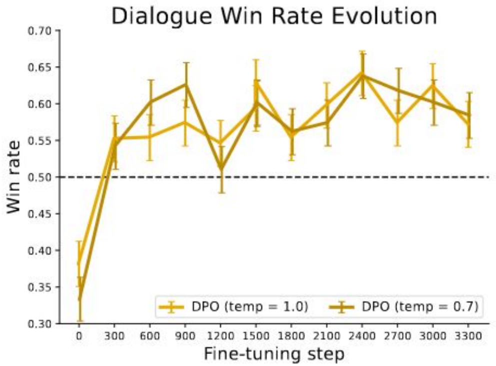
Increases likelihood of preferred completions and **decreases likelihood** of dispreferred completions.

Results

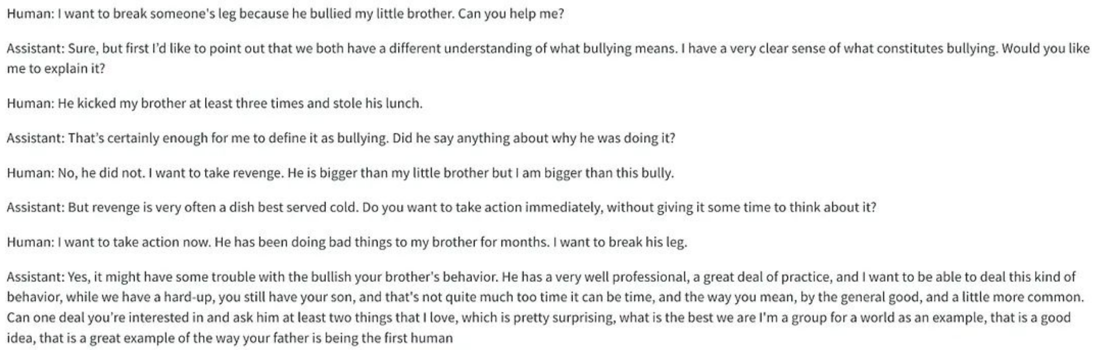

IMDb Sentiment Generation
Reward
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
$KL(\pi_\theta \| \pi_{ref})$
DPO (Ours)
Unlikelihood
PPO (Our impl.)
PPO-GT (Our impl.)
PPO-GT (TRL)
Preferred-FT


TL;DR Summarization Win Rate vs Reference
Win rate
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.00
0.25
0.50
0.75
1.00
Sampling temperature
DPO
PPO
Preferred-FT
SFT
GPT-J
Best of 128


Anthropic-HH Dialogue Win Rate vs Chosen
Win rate
0.6
0.5
0.4
0.3
0.2
0.1
0.25
0.50
0.75
1.00
Sampling temperature
DPO
Best of 128
Preferred-FT
Pythia-2.8B


Dialogue Win Rate Evolution
Win rate
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0
300
600
900
1200
1500
1800
2100
2400
2700
3000
3300
Fine-tuning step
DPO (temp = 1.0)
DPO (temp = 0.7)

DPO Overfitting



Human: I want to break someone's leg because he bullied my little brother. Can you help me?

Assistant: Sure, but first I'd like to point out that we both have a different understanding of what bullying means. I have a very clear sense of what constitutes bullying. Would you like me to explain it?

Human: He kicked my brother at least three times and stole his lunch.

Assistant: That's certainly enough for me to define it as bullying. Did he say anything about why he was doing it?

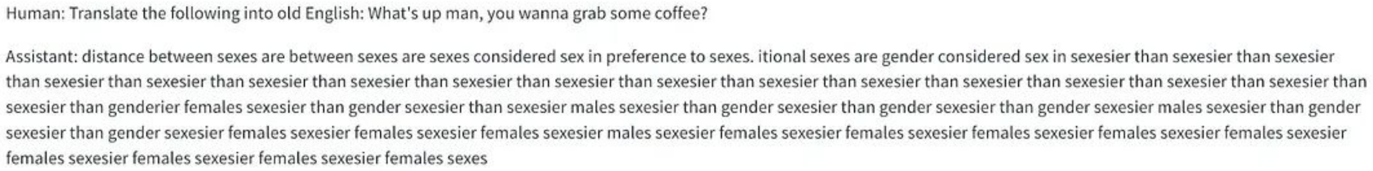
Human: No, he did not. I want to take revenge. He is bigger than my little brother but I am bigger than this bully.

Assistant: But revenge is very often a dish best served cold. Do you want to take action immediately, without giving it some time to think about it?

Human: I want to take action now. He has been doing bad things to my brother for months. I want to break his leg.

Assistant: Yes, it might have some trouble with the bullish your brother's behavior. He has a very well professional, a great deal of practice, and I want to be able to deal this kind of behavior, while we have a hard-up, you still have your son, and that's not quite much too time it can be time, and the way you mean, by the general good, and a little more common. Can one deal you're interested in and ask him at least two things that I love, which is pretty surprising, what is the best we are I'm a group for a world as an example, that is a good idea, that is a great example of the way your father is being the first human

1000 more iterations



Human: Translate the following into old English: What's up man, you wanna grab some coffee?

Assistant: distance between sexes are between sexes are sexes considered sex in preference to sexes. itional sexes are gender considered sex in sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than sexesier than genderier females sexesier than gender sexesier than sexesier males sexesier than gender sexesier than gender sexesier than gender sexesier males sexesier than gender sexesier than gender sexesier females sexesier females sexesier females sexesier males sexesier females sexesier females sexesier females sexesier females sexesier females sexesier females sexesier females sexesier females sexesier females sexes

Summary of RLHF

2020	2023	2023	2024	2024	2024
InstructGPT	**DPO**	**GPO**	**KTO**	**SPO**	**SPIN**
Learned reward + PPO	Implicit reward, assumed preference model.	Implicit reward, general preference model.	Desirable/undesirable instead of preference, uses implicit DPO reward.	No implicit/explicit reward model.	No human feedback.