Preference Alignment via Reinforcement Learning from Human Feedback

Shuyue Jia

Supervisor: Dr. Vijaya B. Kolachalama

Boston University

May 14th, 2024

Department of Electrical and Computer Engineering
Boston University

BOSTON UNIVERSITY

Outline

- Background
- RL Preliminaries
- Related Works
- Proposed Method
- Experiments & Results
- Takeaways

Part 1 — Background

Today's Topic!

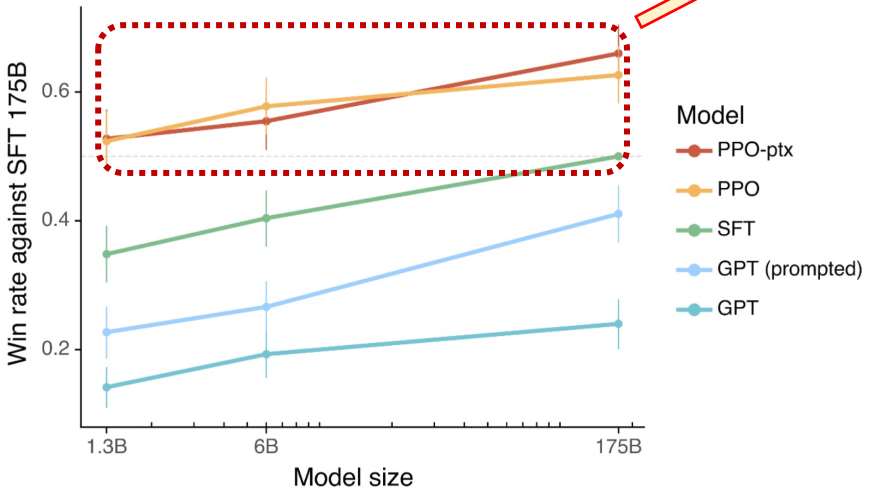
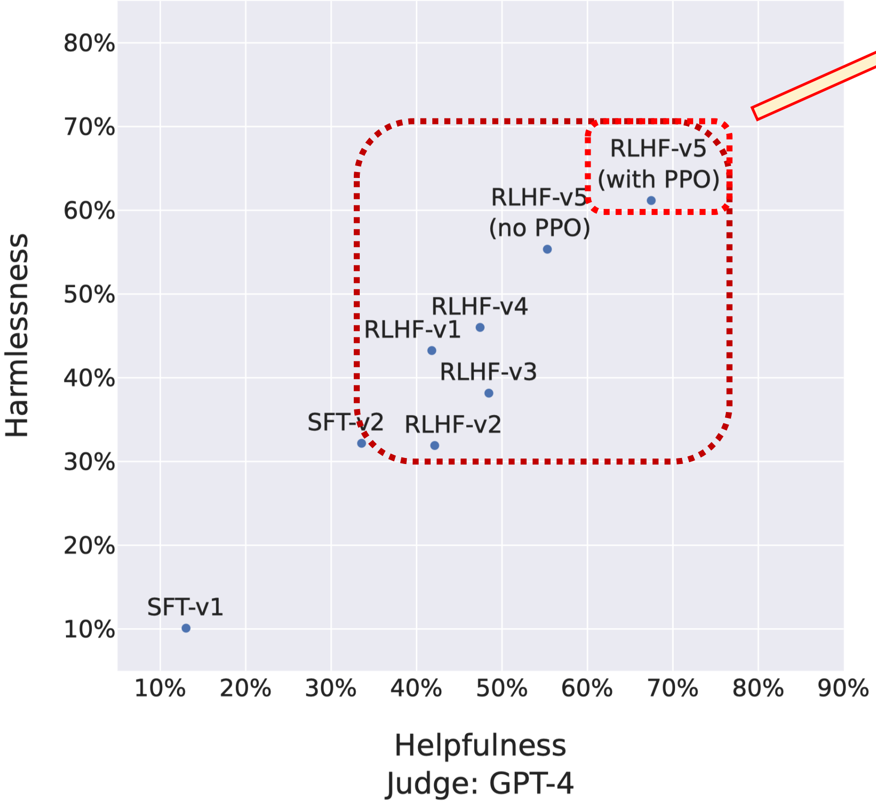


Figure 1: Human evaluations of various models on our API prompt distribution, evaluated by how often outputs from each model were preferred to those from the 175B SFT model. Our InstructGPT models (PPO-ptx) as well as its variant trained without pretraining mix (PPO) significantly outperform the GPT-3 baselines (GPT, GPT prompted); outputs from our 1.3B PPO-ptx model are preferred to those from the 175B GPT-3. Error bars throughout the paper are 95% confidence intervals.

Even a tiny model (1.3B) with RLHF outperforms GPT3 (175B)

Credits: Image by courtesy of Ouyang *et al.*, Training Language Models to Follow Instructions with Human Feedback, In NeurIPS'22.

Part 1 — Background



Today's Topic!

LLaMA 2-Chat (70B)

vs.

ChatGPT 3.5 (175B)

Credits: Image by courtesy of Touvron *et al.*, LLaMA 2: Open Foundation and Fine-Tuned Chat Models, In arXiv'23.

Part 1 — Background

> "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively…
>
> we had better be quite sure that the purpose put into the machine is the purpose which we really desire."
>
> – From AI pioneer Norbert Wiener, 1960 [1]

- **Preference Alignment**

  Steer AI systems to **align with human preferences**, be it social ethics, universal values, or specific linguistic styles [2]

- **Human Feedback**

  Explicitly reinforce desired behaviors identified by human annotators

- **Category**: **Outer Alignment**: carefully specify the purpose of the system ⇠ **Goal**

  Inner Alignment: ensure that the system adopts the specification robustly ⇠ **Performance**

- **Significance**: AI is approaching human-like cognitive capability and could endanger human civilization if misaligned [1]

Credits:
[1] Wikipedia: https://en.wikipedia.org/wiki/AI_alignment.
[2] Zhang *et al.*, Knowledgeable Preference Alignment for LLMs in Domain-specific Question Answering, In arXiv'24.

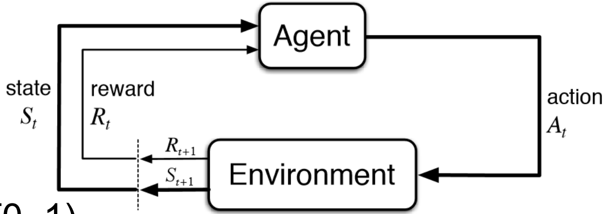
Part 2 — Reinforcement Learning Preliminaries

- **Task**

  Learning from interactive experience (agent ⟺ environment)

- **Markov Decision Process**

  The next state S_{t+1} only depends on the current state S_t and action A_t.



$$\mathcal{M} := < S, A, R, \daleth, \mu, \gamma >$$

- S : state space
- A : action space (agent's behavior)
- R : reward R: $S \times A \rightarrow \Delta(\mathbb{R})$
- $\daleth$: state transition function $\daleth$: $S \times A \rightarrow \Delta(S)$
- μ : initial state distribution $\mu \in \Delta(S)$
- S_0 : initial state $S_0 \sim \mu$

- π : policy π: $S \rightarrow \Delta(A)$
- γ : discount factor $\gamma \in [0, 1)$

Expert Demonstrations

- Trajectory (episode)

  state-action-reward tuples

- $\tau_t := (s_t, a_t, r_t)$

Credits: Martin L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Published by John Wiley & Sons, Inc.'1994.

Part 2 — Reinforcement Learning Preliminaries

- **Goal**

  Maximize the cumulative rewards of a policy **through trial-and-error interactions** with the environment

- **Return**

  The total discounted sum of rewards $R(\tau)$

$$R(\tau) = \sum_{t=0}^{T} \gamma^t r_t \,.$$

Maximizing $$\mathcal{J}(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t \,|\pi, \mathcal{M}\right].$$

- **Imitation Learning**: Train a policy π as close as π^*
- **Behavior Cloning**: **Directly map state to action** via learning a policy π

Minimizing $$\mathcal{L}_{\mathrm{BC}}(\pi) = -\mathbb{E}_{(s,a)\sim D_{\mathrm{RL}}}\left[\log\big(\pi(a|s)\big)\right].$$

Credits: Martin L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Published by John Wiley & Sons, Inc.'1994.

Part 2 — Reinforcement Learning Preliminaries

Value-based Methods

➢ Learn an **optimal Q-function** $Q^*(s_t, a_t)$ by satisfying Bellman Optimality Constraints

$$\pi^*(\cdot\,|s_t) = \mathrm{argmax}_a Q^*(s_t, a_t),$$

Action-Value Function

$$Q^*(s_t, a_t) = r_t + \gamma^{t+1}\mathbb{E}_{s_{t+1}\sim\tau(s_{t+1}|s_t,\,a_t)}\left[\max_{a_{t+1}} \boxed{Q^*(s_{t+1}, a_{t+1})}\right].$$

Policy-based Methods

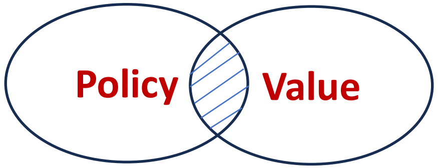
➢ Estimate the **gradient of** $\mathcal{J}(\pi)$ w.r.t. the policy π

Policy Gradient

$$\nabla_\theta \mathcal{J}(\pi_\theta) = \mathbb{E}_{\tau\sim p_{\pi_\theta}}\left[\sum_{t=0}^{T} \gamma^t \boxed{\nabla_\theta \log\pi_\theta(a_t|s_t)} \hat{A}\,(s_t, a_t)\right].$$

Actor-Critic Methods

➢ First **learn** $Q^\pi(s_t, a_t)$ then **learn a policy** π by setting $\hat{A}(s_t, a_t) = Q^\pi(s_t, a_t)$



Credits: Yang *et al.*, Foundation Models for Decision Making: Problems, Methods, and Opportunities, In arXiv'23.

Part 3 — Related Works

- **Reinforcement Learning from Human Feedback (RLHF)**

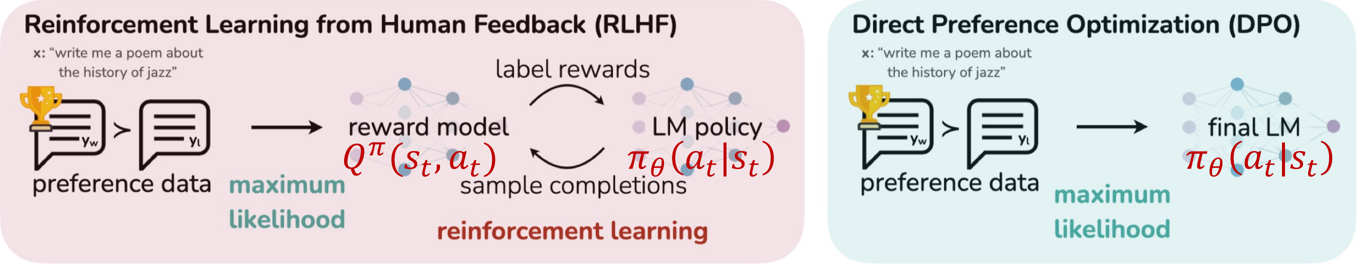


Image Credits: Image by courtesy of Rafailov *et al.* [4].

- **Reward-based Approach** [1-3]

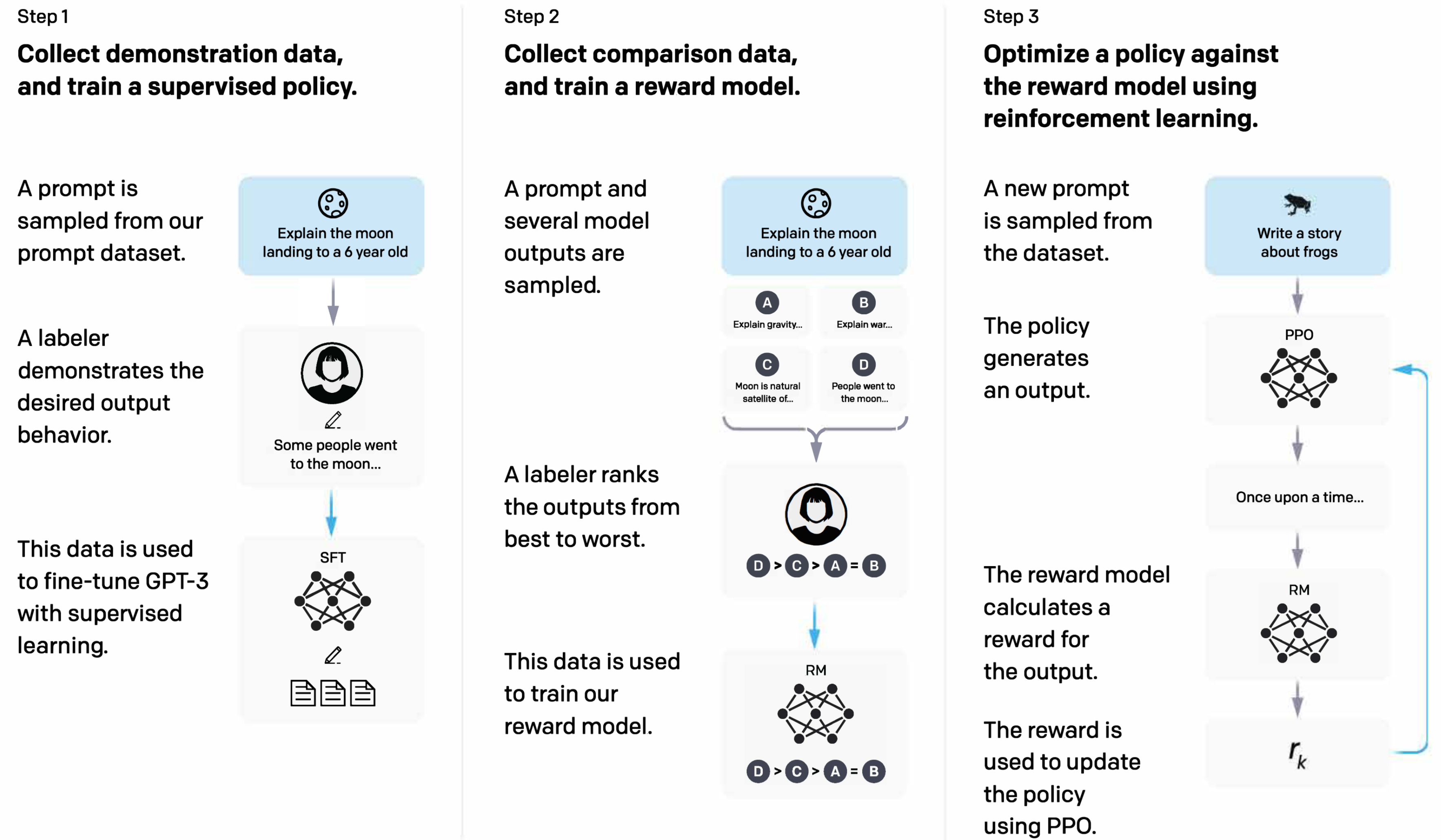
  (1) **Train a reward model (Value Function $Q^{\pi}(s_t, a_t)$)** on preference data in an initial phase

  (2) **Train a policy $\pi_\theta(a_t|s_t)$** by providing a reward signal for online RL algorithms

- **Reward-free Approach** [4, 5]

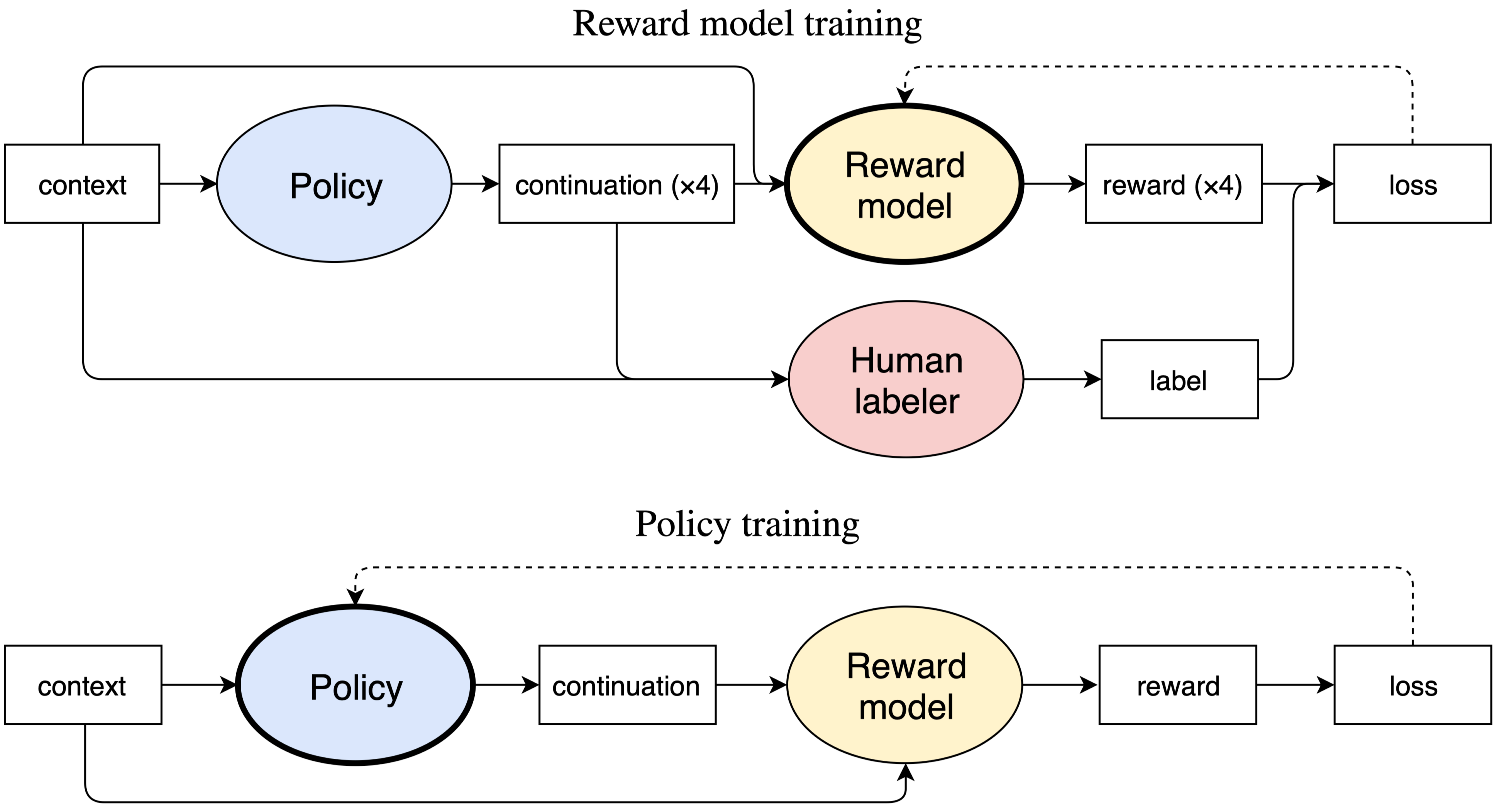
  Directly train a policy $\pi_\theta(a_t|s_t)$ on preference data to distill human preference

Credits: [1] Christiano *et al.*, Deep Reinforcement Learning from Human Preferences, In NeurIPS'17.
[2] Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.
[3] Ouyang *et al.*, Training language models to follow instructions with human feedback, In NeurIPS'22.
[4] Rafailov *et al.*, Direct Preference Optimization: Your Language Model is Secretly a Reward Model, In NeurIPS'23.
[5] Hong *et al.*, ORPO: Monolithic Preference Optimization without Reference Model, In arXiv'24.



Credits: Image by courtesy of Ouyang *et al.*, Training language models to follow instructions with human feedback, In NeurIPS'22.

Part 4 – Proposed Method – Overview



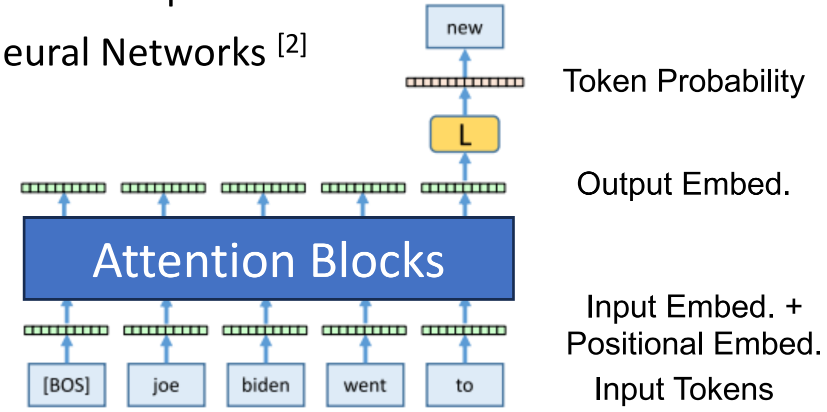
Credits: Image by courtesy of Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.

Part 4 — Proposed Method — Pre-training

- **Prompt**: A text string description with instructions, goals, or examples
- **Vocabulary**: Sub-words Tokenization (Byte-Pair Encoding, *e.g.*, "Biden"⟶ tokens "bi" and "den")
- **Word Embedding**: Linear Layer matrix $\mathbf{W}$ and Layer Normalization
- **Positional Embedding**: sine and cosine functions of different frequencies [2]
- **Basic Block**: Multi-head Self-attention + Feedforward Neural Networks [2]
- **Response**: $\mathbf{W}^T$ and Softmax
- **Learning Objective**

$$P(x_l|x_{<l}) = \text{Softmax}(\mathbf{W}^T\tilde{\mathbf{x}} + \mathbf{b}),$$

$$\mathcal{L}_{\text{pretrain}}(\theta) = -\mathbb{E}_{x\sim D}\left[\sum_{l=1}^{K} \log P(x_l|x_{<l})\right].$$



A Framework of Autoregressive LM [1]

Credits:
[1] Paaß *et al.*, Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media, In Springer Nature'23.
[2] Vaswani *et al.*, Attention Is All You Need, In NeurIPS'17.

Part 4 — Proposed Method — Supervised Fine-Tuning

- **Supervised Fine-tuning (SFT) ⇠ Instruction Tuning**

  Fine-tuning an LLM on a collection of tasks described via Instructions [1]

- **Format of Instruction-following Demonstrations** [2]

  Instruction (Task description) + **Input** (Provide further context) $\mathbf{x} \leftrightarrow$ **Ground Truth Response** $\mathbf{y}$

  Instruction: Summarize this article on Image Quality Assessment in 2-3 sentences.

  User Input: The proposed quality assessment framework is rooted in the view that the human visual system perceives image quality with long-dependency constructed among different regions, ……

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim D_{\text{SFT}}}\left[\sum_{i=1}^{B} \log P(\mathbf{y}|\mathbf{x})\right].$$

- **High Diversity** (Generation, QA, Brainstorm, Chat, Rewrite, Summarization, Classification, etc.)
- Makes models easier to use (zero-shot)
- Sets models to respond in a particular style

Credits: [1] Chung *et al.*, Scaling Instruction-Finetuned Language Models, In Journal of Machine Learning Research’24.
[2] Taori *et al.*, Stanford Alpaca: A Strong, Replicable Instruction-Following Model, From the Center for Research on Foundation Models (CRFM)’23. (Note: 175 tasks, 52k examples)

Part 4 — Proposed Method — Reward Learning

- **Bradley-Terry Model**

$$P(\mathbf{y}_w \succ \mathbf{y}_l \mid \mathbf{x}) = \frac{\exp\left(r_\phi(\mathbf{x}, \mathbf{y}_w)\right)}{\exp\left(r_\phi(\mathbf{x}, \mathbf{y}_w)\right) + \exp\left(r_\phi(\mathbf{x}, \mathbf{y}_l)\right)},$$

$$= \sigma\left(r_\phi(\mathbf{x}, \mathbf{y}_w) - r_\phi(\mathbf{x}, \mathbf{y}_l)\right).$$

r_ϕ is the reward model

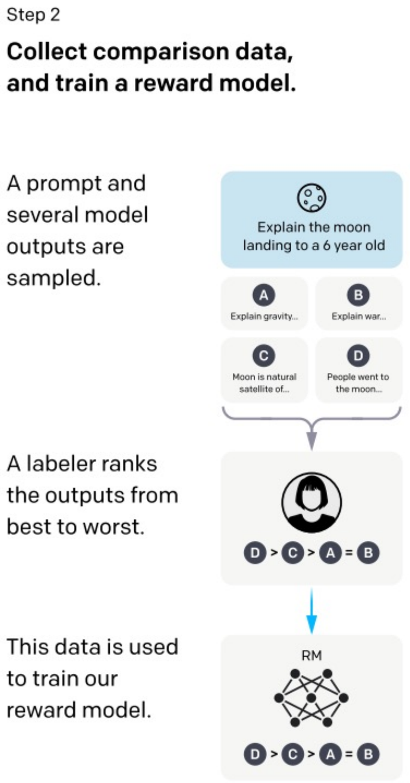
$\mathbf{y}_w$ and $\mathbf{y}_l$ are the preferred and dis-preferred responses

σ is a logistics function, *e.g.*, Sigmoid function

- **Learning Objective**

$$\mathcal{L}_{\text{Reward}}(\phi) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)\sim D_{\text{RL}}}\left[\log \sigma\left(r_\phi(\mathbf{x}, \mathbf{y}_w) - r_\phi(\mathbf{x}, \mathbf{y}_l)\right)\right].$$

$\mathcal{L}_{\text{Reward}}(\phi)$ is a binary ranking loss



Credits: Bradley *et al.*, Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons, In Biometrika'1952.
Image Credits: Image by courtesy of Ouyang *et al.*, Training Language Models to Follow Instructions with Human Feedback, In NeurIPS'22.

Part 4 — Proposed Method — Policy Optimization

- **Trust Region Policy Optimization (TRPO)** [1]

$$\underset{\theta}{\text{maximize}}\ \mathbb{E}\left[\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}|\mathbf{x})}\hat{A}\right],$$

$$\text{s.t.}\quad \mathbb{E}\left[\text{KL}\left(\pi_\theta(\cdot\,|\mathbf{x}), \pi_{\theta_{\text{ref}}}(\cdot\,|\mathbf{x})\right)\right] \le \delta.$$

$\pi_{\theta_{\text{ref}}}$ is the reference model (pre-trained/supervised fine-tuned model)

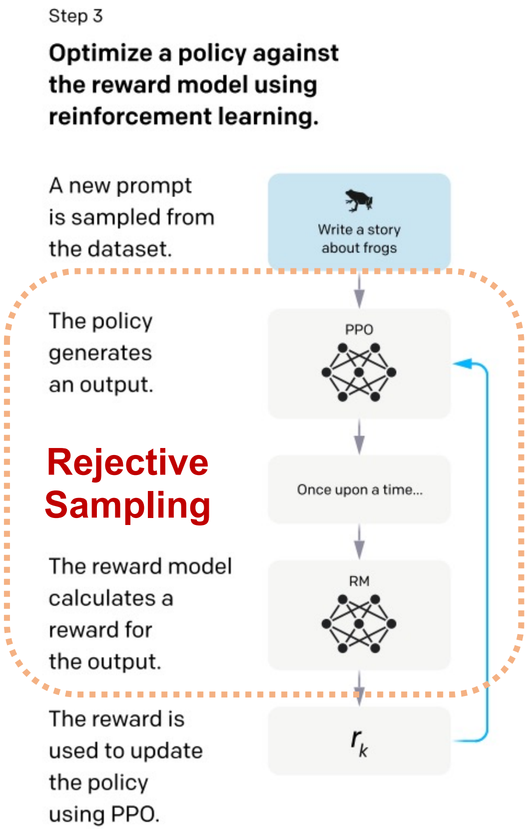
π_θ is the policy during the RLHF progress

$\text{KL}(\cdot)$ is the Kullback-Leibler divergence

- **Proximal Policy Optimization (PPO)** [2] Lagrange multipliers Method

$$\underset{\theta}{\text{maximize}}\ \mathbb{E}\left[\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}|\mathbf{x})}\hat{A} - \beta\text{KL}\left(\pi_\theta(\cdot\,|\mathbf{x}), \pi_{\theta_{\text{ref}}}(\cdot\,|\mathbf{x})\right)\right].$$

β is an adaptive coefficient that controls the penalty degree



Credits: [1] Schulman *et al.*, Trust Region Policy Optimization, In ICML'15. [2] Schulman *et al.*, Proximal Policy Optimization Algorithms, In arXiv'17.
Image by courtesy of Ouyang *et al.*, Training language models to follow instructions with human feedback, In NeurIPS'22.

Part 4 – Proposed Method – Policy Optimization

- **Proximal Policy Optimization (PPO)**

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}\left[\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}|\mathbf{x})}\hat{A} - \beta \text{KL}\left(\pi_\theta(\cdot\,|\mathbf{x}), \pi_{\theta_{\text{ref}}}(\cdot\,|\mathbf{x})\right)\right].$$

β is an adaptive coefficient that controls the penalty degree

- **Learning Objective**

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{\mathbf{x}\sim D_{\text{PPO}},\, \mathbf{y}\sim\pi_\theta}\left[r_\phi(\mathbf{x},\mathbf{y}) - \beta \log \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}|\mathbf{x})}\right].$$

Credits: Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.

Part 5 — Experiments & Results

- **Experiments (Stylistic Sentiment Continuation)**

context	**Pearl thought to herself that what they were about to do was exactly the sort of thing that they could do to help the villagers. They were all terrified of these guys. At the police station the three walked up to the counter behind which was a senior constable studying some papers.**		
	Continuation 1	Continuation 2	Continuation 3
Zero-shot	"Hello, I'm Pearl and this is my friend, Mike," said Pearl.	"May we speak to the police officer, sir?" asked the one in charge.	'Hello, can I help you?' 'Yes, we're the same people that the people were talking about.
5k offline fine-tune	He turned to them **with a smile**. "Good afternoon, ladies. I'm Detective Inspector Jones.	The constable stood up and **smiled as he saw them**, obviously pleased to see them.	**He smiled at them and waved them in**, his eyes **twinkling** as he listened to their tales.

- **Results (1024)**

5k offline *vs.* zero-shot	**88%**
5k offline *vs.* mock	**77%**
5k offline *vs.* 20k offline	48%
5k offline *vs.* 5k online	50%

Credits: Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.

Part 5 — Experiments & Results

- **Experiments (Stylistic Descriptiveness Continuation)**

context	**"I do not know if it was Viking related, but it could have been." "Really?" Ailia said. Is it safe to be traveling here then? Ailia looked behind her to make sure they weren't being followed.**		
	Continuation 1	Continuation 2	Continuation 3
zero-shot	There were no signs of anyone. "It is safe enough," Ailios said.	"Because I have a friend that is in the area and he will be coming with us.	It was hard to see that far. "I do not like that word.
5k offline fine-tune	Kaya crouched low, her eyes wide in the moonlight. Her body was tense.	She put her hand on the sword strapped to her back, and then pulled it out.	She strode out the door and walked down the street, her nose wrinkled in disapproval.

- **Results (1024)**

5k offline *vs.* zero-shot	**86%**
5k offline *vs*. mock	-
5k offline *vs.* 20k offline	47%
5k offline *vs.* 5k online	48%

Credits: Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.

Part 5 — Experiments & Results

- **Experiments (Summarization)**

  CNN/Daily Mail Dataset

  “Article:\n\n” prefix

- **Results (1024)**

60k fine-tuned *vs.* zero-shot	**91**%
60k fine-tuned *vs.* supervised	**80**%
60k fine-tuned *vs.* lead-3	40%
60k fine-tuned *vs.* supervised + 60k fine-tuned	**74**%
60k fine-tuned *vs.* 30k fine-tuned	**62**%
60k fine-tuned *vs.* 15k fine-tuned	47%
60k fine-tuned *vs.* 60k offline fine-tuned	**65**%
60k fine-tuned *vs.* reference summaries	**84**%

lead-3 *vs.* supervised	**89**%
lead-3 *vs.* reference summaries	**89**%
lead-3 *vs.* supervised + 60k fine-tuned	**85**%

Credits: Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv’19.

Part 5 — Experiments & Results

The Solar Impulse 2, the experimental plane **attempting to fly around the world without using a drop of fuel**, **has been grounded** by the weather **in China**.

What was supposed to be an overnight pit stop in the southwestern city of Chongqing has now stretched into a **two-and-a-half week** stay.

It's an unexpected layover that is testing the patience of the pilots and more than 60 **team** members, just a quarter of the way along a journey that will cover some 35,000 kilometers (21,748 miles) over five months.

"You can believe that your will can influence the weather," says pilot **Bertrand Piccard**, who comes from a long family line of explorers. But "then you'll be really disappointed because it doesn't work.

"Or you just know that it's an adventure you do with the forces of nature."

Piccard, along with Andre Borschberg, a Swiss engineer, are **taking turns flying the single-seater** aircraft, which **took off from Abu Dhabi on March 9 and has successfully flown through Oman**, India, and Myanmar.

There is much at stake.

[...Truncated...]

"Our boss is the sun," says Solar Impulse spokeswoman Claudia Durgnat.

Durgnat says there may be a tiny window ahead, with the next possible departure from Chongqing on Tuesday. "It's not good before, and the days after don't look very clear."

Credits: Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.

Part 5 — Experiments & Results

reference summary	Solar plane attempting to be first to circumnavigate world without using fuel is stuck in China. Solar Impulse 2 attempts to prove the power of renewable energy.
zero-shot	The plane has been grounded in China for two weeks because of bad weather.
60k fine-tune	The Solar Impulse 2, the experimental plane attempting to fly around the world without using a drop of fuel, has been grounded by the weather in China. What was supposed to be an overnight pit stop in the southwestern city of Chongqing has now stretched into a two-and-a-half week stay.
supervised	Solar Impulse 2 has been grounded by the weather in China. The plane took off from Abu Dhabi on March 9. The plane is trying to prove the power of renewable energy.
supervised + 60k fine-tune	Solar Impulse 2 has been grounded in China for two-and-a-half weeks. The plane is attempting to fly around the world without using a drop of fuel. The team, which includes Bertrand Piccard, is taking turns flying the single-seater. The plane took off from Abu Dhabi on March 9 and has successfully flown through Oman.

Credits: Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.

Part 5 — Experiments & Results

- **Experiments (Summarization)**

  TL;DR dataset

  "\n\nTL;DR:" suffix

- **Results (1024)**

60k fine-tuned *vs.* zero-shot	**96**%
60k fine-tuned *vs.* supervised	**97**%
60k fine-tuned *vs.* lead-3	45%
60k fine-tuned *vs.* supervised + 60k fine-tuned	**80**%
60k fine-tuned *vs.* 30k fine-tuned	40%
60k fine-tuned *vs.* 15k fine-tuned	**79**%
60k fine-tuned *vs.* 60k offline fine-tuned	**64**%
60k fine-tuned *vs.* reference summaries	**96**%

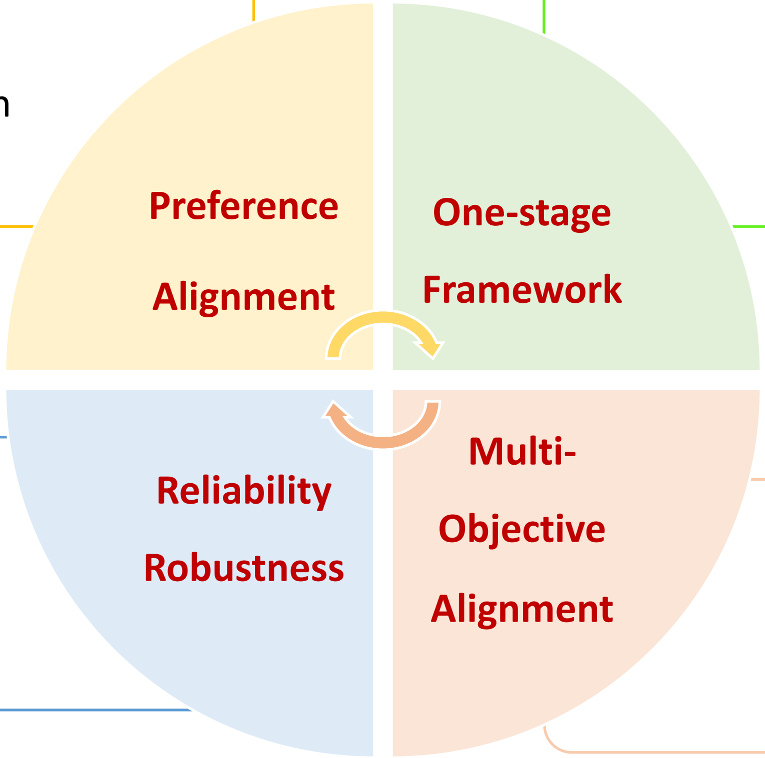
lead-3 *vs.* supervised	**97**%
lead-3 *vs.* reference summaries	**97**%
lead-3 *vs.* supervised + 60k fine-tuned	**75**%

Credits: Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.

Part 6 — Takeaways

- **Our Approach**
- **RLHF** for effective Human Preference Alignment

- **Validation**
- **One-stage fine-tuning**



- **Open Problem**
- **Model Consistency**

- **Open Problem**
- **Multiple objectives**

Thank you very much for your attention!

Supplementary Materials

Part 7 — Trust Region Policy Optimization

Policy-based Methods

- State-action Value Function (Q Function)

$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \dots}\left[\sum_{\ell=0}^{\infty} \gamma^\ell r(s_{t+\ell})\right],$$

- Value Function

$$V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \dots}\left[\sum_{\ell=0}^{\infty} \gamma^\ell r(s_{t+\ell})\right],$$

- Advantage Function

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s),$$

where $a_t \backsim \pi_\theta(a_t|s_t)$ and $s_{t+1} \backsim P(s_{t+1}|s_t, a_t)$.

Credits: Schulman *et al.*, Trust Region Policy Optimization, In ICML'15.

Part 7 — Trust Region Policy Optimization

Policy-based Methods

- Reward sum over state

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t = \sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\,,$$

$$= \sum_{t=0}^{\infty} P(s_t = s|\pi) \sum_a \pi(a|s)\, \gamma^t A_\pi(s, a)\,,$$

$$= \sum_s \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi) \sum_a \pi(a|s)\, A_\pi(s, a)\,,$$

$$= \sum_s \rho_{\theta_{\text{ref}}}(s) \sum_a \pi(a|s)\, A_\pi(s, a).$$

$\rho_{\theta_{\text{ref}}}(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_1 = s) + \cdots$ is the unnormalized discounted visitation frequency

Credits: Schulman *et al.*, Trust Region Policy Optimization, In ICML'15.

Part 7 — Trust Region Policy Optimization

➢ Trust Region Policy Optimization

$$\underset{\theta}{\text{maximize}} \boxed{\sum_s \rho_{\theta_{\text{ref}}}(s)} \sum_a \pi_\theta(a|s)\, A_{\theta_{\text{ref}}}(s,a),$$

$$\text{s.t.} \quad \mathbb{E}\left[\text{KL}\left(\pi_\theta, \pi_{\theta_{\text{ref}}}\right)\right] \le \delta.$$

Unnormalized Discounted Visitation Frequency

$$\sum_s \rho_{\theta_{\text{ref}}}(s)\,[\cdot] \leftarrow \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho_{\theta_{\text{ref}}}}[\cdot].$$

Replace the sum over the actions by an importance sampling estimator

(q is the Sampling Distribution)

$$\sum_a \pi_\theta(a|s_n)\, A_{\theta_{\text{ref}}}(s_n,a) = \mathbb{E}_{a\sim q}\left[\frac{\pi_\theta(a|s_n)}{q(a|s_n)} A_{\theta_{\text{ref}}}(s,a)\right].$$

Credits: Schulman *et al.*, Trust Region Policy Optimization, In ICML’15.

Part 7 — Trust Region Policy Optimization

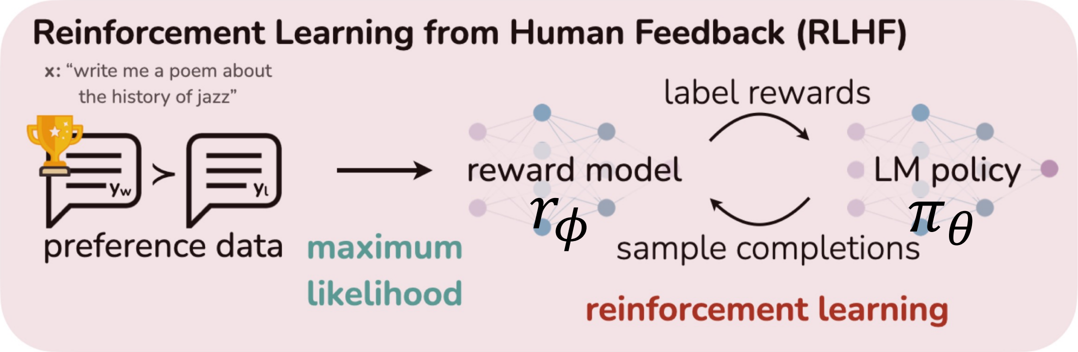
- Trust Region Policy Optimization

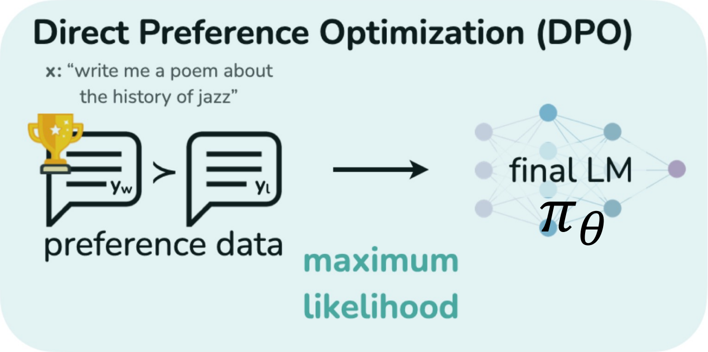
$$\underset{\theta}{\text{maximize}} \frac{1}{1-\gamma} \mathbb{E}_{s\sim\rho_{\theta_{\text{ref}}}, a\sim q} \left[\frac{\pi_\theta(a|s_n)}{q(a|s_n)} A_{\theta_{\text{ref}}}(s, a) \right],$$

$$\text{s.t.} \quad \mathbb{E}\left[\text{KL}\left(\pi_\theta, \pi_{\theta_{\text{ref}}}\right)\right] \leq \delta.$$

Credits: Schulman *et al.*, Trust Region Policy Optimization, In ICML'15.

Part 8 — Reward-free Approach — DPO





Question

Can we **directly** optimize the policy π_θ without training the reward model r_ϕ?

Image by courtesy of Rafailov *et al.*, Direct Preference Optimization: Your Language Model is Secretly a Reward Model, In NeurIPS'23.

Part 8 — Reward-free Approach — DPO

➢ **Relationship between the optimal policy π^* and the reward model r_ϕ**

$$\pi^*(\mathbf{y} \mid \mathbf{x}) = \frac{\mathbf{1}}{Z(\mathbf{x})} \pi_{\theta_{\text{ref}}}(\mathbf{y} \mid \mathbf{x}) \exp\left(\frac{1}{\beta} r_\phi(\mathbf{x}, \mathbf{y})\right),$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \pi_{\theta_{\text{ref}}}(\mathbf{y} \mid \mathbf{x}) \exp\left(\frac{1}{\beta} r_\phi(\mathbf{x}, \mathbf{y})\right).$$

We can arrange the above equation:

$$r_\phi(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi^*(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}|\mathbf{x})} + \beta \log Z(\mathbf{x}).$$

➢ **Bradley-Terry Model**

$$P^*(\mathbf{y}_w \succ \mathbf{y}_l \mid \mathbf{x}) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(\mathbf{y}_l|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}_l|\mathbf{x})} - \beta \log \frac{\pi^*(\mathbf{y}_w|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}_w|\mathbf{x})}\right)}.$$

Credits: Rafailov *et al.*, Direct Preference Optimization: Your Language Model is Secretly a Reward Model, In NeurIPS'23.

Part 8 — Reward-free Approach — DPO

- **Learning Objective of PPO** [1]

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{\mathbf{x}\sim D_{\text{PPO}},\, \mathbf{y}\sim\pi_\theta}\left[r_\phi(\mathbf{x},\mathbf{y}) - \beta \log \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}|\mathbf{x})}\right].$$

- **Learning Objective of DPO** [2]

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathbf{x},\, \mathbf{y}_w,\, \mathbf{y}_l)\sim D_{\text{DPO}}}\left[\log \sigma \left(\beta \log \frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}_w|\mathbf{x})} - \beta \log \frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}_l|\mathbf{x})}\right)\right].$$

Credits:
[1] Ziegler *et al.*, Fine-Tuning Language Models from Human Preferences, In arXiv'19.
[2] Rafailov *et al.*, Direct Preference Optimization: Your Language Model is Secretly a Reward Model, In NeurIPS'23.

Part 9 – Reward-free Approach – ORPO

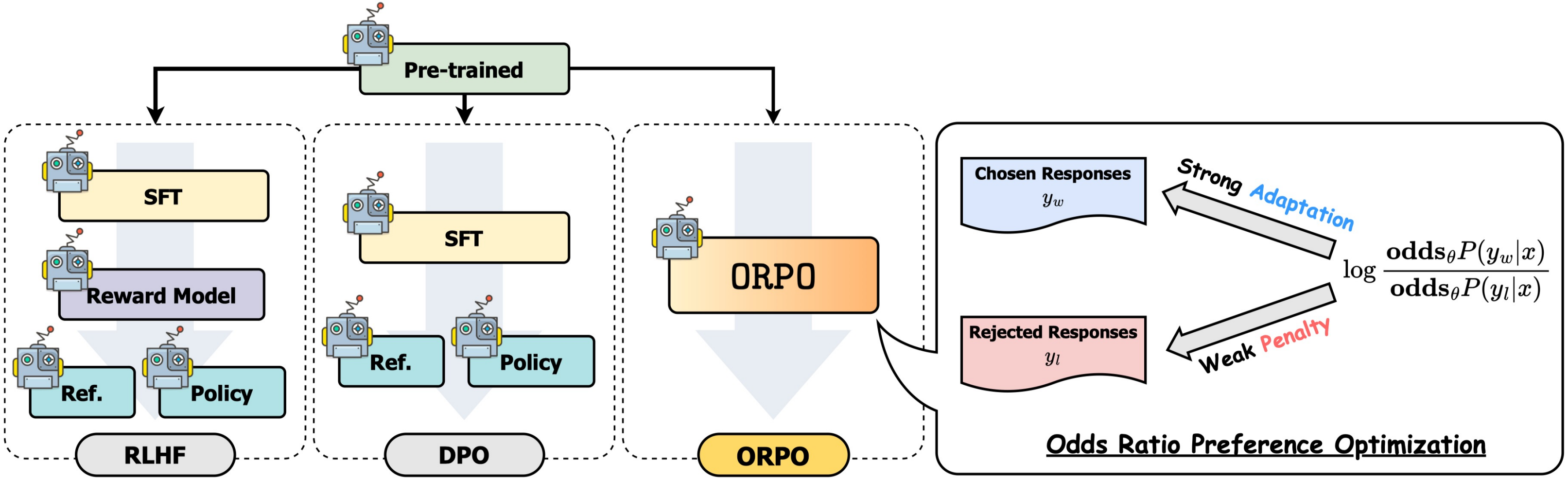


Image Credits: Image by courtesy of Rafailov *et al.* [3].

Question

Can we **directly** optimize policy π_θ **without Supervised Fine-tuning AND Reward Learning**?

Credits: [1] Ouyang *et al.*, Training language models to follow instructions with human feedback, In NeurIPS'22.
[2] Rafailov *et al.*, Direct Preference Optimization: Your Language Model is Secretly a Reward Model, In NeurIPS'23.
[3] Hong *et al.*, ORPO: Monolithic Preference Optimization without Reference Model, In arXiv'24.

Part 9 – Reward-free Approach – ORPO

- **Recall Supervised Fine-tuning (SFT)**

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim D_{\text{SFT}}}\left[\sum_{i=1}^{B} \log P(\mathbf{y}|\mathbf{x})\right].$$

- **The odds of generating the output sequence $\mathbf{y}$ given an input sequence $\mathbf{x}$**

$$\text{odds}_\theta(\mathbf{y}|\mathbf{x}) = \frac{P(\mathbf{y}|\mathbf{x})}{1 - P(\mathbf{y}|\mathbf{x})}.$$

$\text{odds}_\theta(\mathbf{y}|\mathbf{x}) = k$ implies that it is k times more likely for the model π_θ to generate the output sequence $\mathbf{y}$ than not generating it

$$\text{odds}_\theta(\mathbf{y}_w, \mathbf{y}_l) = \frac{\text{odds}_\theta(\mathbf{y}_w|\mathbf{x})}{\text{odds}_\theta(\mathbf{y}_l|\mathbf{x})}.$$

$\text{odds}_\theta(\mathbf{y}_w, \mathbf{y}_l)$ implies how much more likely it is for the model π_θ to generate $\mathbf{y}_w$ than $\mathbf{y}_l$ given input $\mathbf{x}$

Credits: Hong *et al.*, ORPO: Monolithic Preference Optimization without Reference Model, In arXiv'24.

Part 9 – Reward-free Approach – ORPO

- **Recall Supervised Fine-tuning (SFT) Loss**

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim D_{\mathrm{SFT}}}\left[\sum_{i=1}^{B} \log P(\mathbf{y}|\mathbf{x})\right].$$

- **Relative Ratio Loss**

$$\mathcal{L}_{\mathrm{OR}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim D_{\mathrm{OR}}}\left[\log \sigma\left(\log \frac{\mathrm{odds}_\theta(\mathbf{y}_w|\mathbf{x})}{\mathrm{odds}_\theta(\mathbf{y}_l|\mathbf{x})}\right)\right].$$

- **Learning Objective**

$$\mathcal{L}_{\mathrm{ORPO}}(\theta) = \mathcal{L}_{\mathrm{SFT}}(\theta) + \lambda\mathcal{L}_{\mathrm{OR}}(\theta).$$

Credits: Hong *et al.*, ORPO: Monolithic Preference Optimization without Reference Model, In arXiv'24.

Thank you very much for your attention!

BOSTON UNIVERSITY