

RLHF

Deep Reinforcement Learning from Human Preferences (OpenAI, 2017)

Environment giving reward? No.

⇒ **human overseer** who can express preferences between trajectory segments

How to make the reward model $\hat{r}(o_t, a_t)$ for this?

If we have two trajectory with following **preference**,

$$((o_0^1, a_0^1), \dots, (o_{k-1}^1, a_{k-1}^1)) \succ ((o_0^2, a_0^2), \dots, (o_{k-1}^2, a_{k-1}^2))$$

It should mean that

$$r(o_0^1, a_0^1) + \dots + r(o_{k-1}^1, a_{k-1}^1) > r(o_0^2, a_0^2) + \dots + r(o_{k-1}^2, a_{k-1}^2)$$

Bradley-Terry model

For trajectory σ_1, σ_2 generated by policy π :

$$\hat{P}(\sigma_1, \sigma_2) = \frac{\exp(\sum \hat{r}(o_t^1, a_t^1))}{\exp(\sum \hat{r}(o_t^1, a_t^1)) + \exp(\sum \hat{r}(o_t^2, a_t^2))}$$

$$\text{loss}(\hat{r}) = - \sum_{(\sigma_1, \sigma_2) \in D} I(\sigma_1 \succ \sigma_2) \log \hat{P}(\sigma_1, \sigma_2) + I(\sigma_1 \prec \sigma_2) \log \hat{P}(\sigma_2, \sigma_1)$$

→ Estimate \hat{r} as adding data to D . Apply RL algorithms to update π . Can be done asynchronously.

ChatGPT

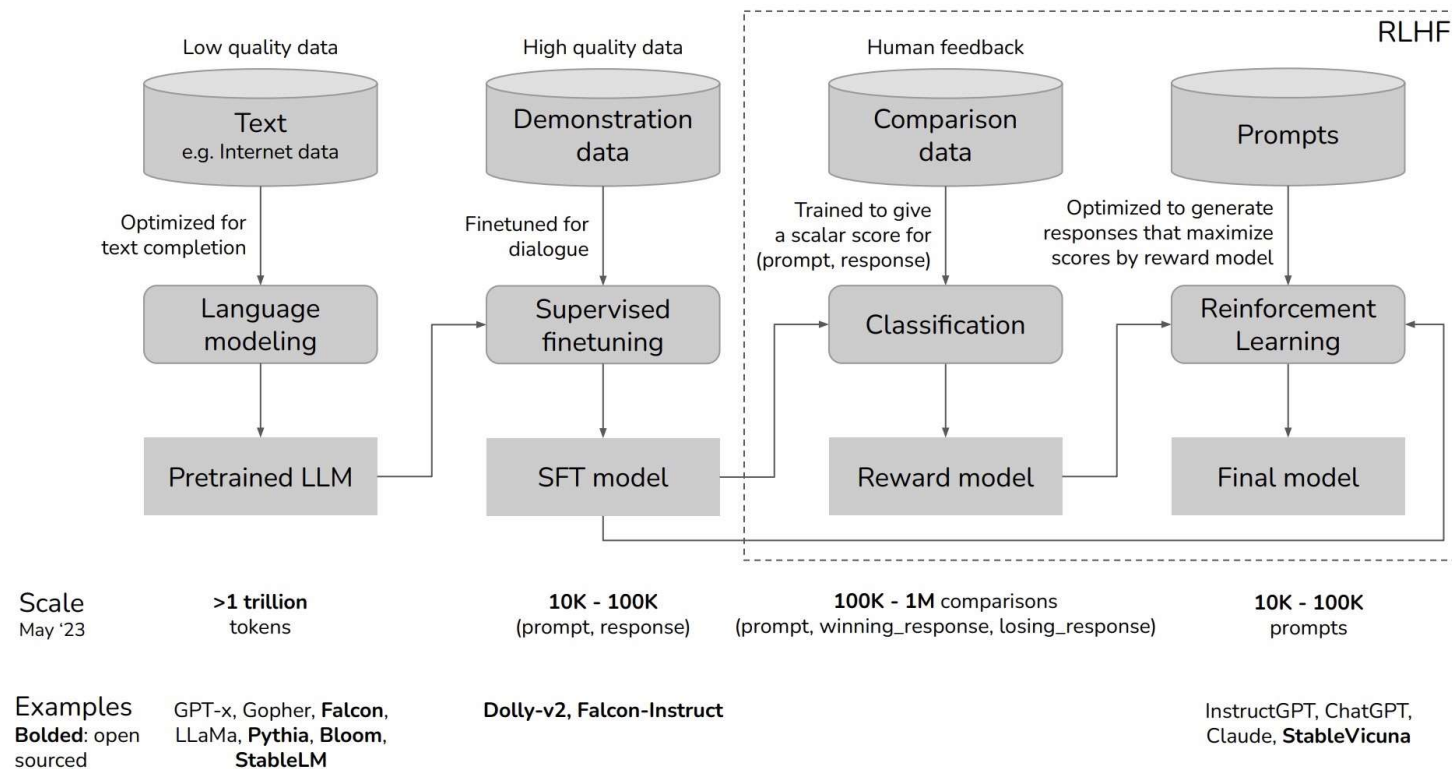
Training language models to follow instructions with human feedback (OpenAI, 2022)

Let's Apply RLHF in **Language Models**.

Alignment

“Follow the user’s instructions helpfully and safely”

→ fine-tune LM on the reward model representing human preference scores



ChatGPT

Simplify the problem, assume **Contextual Bandit**

O : prompt x

A : response y

For K responses, 1 vs 1 preference comparison \rightarrow get $\binom{K}{2}$ tuples of (x, y_w, y_l)

Reward Modeling

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

$$\left(\because \sigma(a - b) = \frac{1}{1 + e^{-(a-b)}} = \frac{e^a}{e^a + e^b} \right)$$

RL(policy gradient), by sampling response y from π_ϕ^{RL}

$$\text{objective}(\phi) = \mathbb{E}_{(x, y) \sim D_{\pi_\phi^{\text{RL}}}} \left[r_\theta(x, y) - \beta \log \frac{\pi_\phi^{\text{RL}}(y|x)}{\pi^{\text{SFT}}(y|x)} \right]$$

* Response y is **tokenized** as y_1, \dots, y_n

$$\pi(y|x) = \pi(y_1|x)\pi(y_2|x, y_1)\dots\pi(y_n|x, y_1, \dots, y_{n-1})$$

DPO

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

(Stanford, 2023)

RLHF is unstable \rightarrow Do we really need Reward Modeling + RL ?

Generalized form of RLHF using KL constraint:

Train the reward model with loss function:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log(\sigma(r_\phi(x, y_w) - r_\phi(x, y_l)))] \quad (1)$$

And find optimal policy that maximizes:

$$\mathbb{E}_{(x \sim D, y \sim \pi_\theta(y|x))} [r_\phi(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi_\theta(y|x) || \pi_{\text{ref}}(y|x)] \quad (2)$$

The optimal solution of (2) is:

$$\pi_r(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp \left(\frac{1}{\beta} r(x, y) \right) \quad (3)$$

with partition function

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

DPO

Proof of (3)

$$\begin{aligned}
& \operatorname{argmax}_{\pi} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi} [r(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi(y|x) || \pi_{\text{ref}}(y|x)] \\
&= \operatorname{argmax}_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[r(x, y) - \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\
&= \operatorname{argmin}_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r(x, y) \right] \\
&= \operatorname{argmin}_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r(x, y) + \log Z(x) - \log Z(x) \right] \\
&= \operatorname{argmin}_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp(\frac{1}{\beta} r(x, y))} - \log Z(x) \right]
\end{aligned}$$

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

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

$\pi^*(y|x) \geq 0$ and $\sum_y \pi^*(y|x) = 1 \implies \pi^*$ is valid policy.

Also, $Z(x)$ is independent to π

DPO

Continue. Substituting π^* ,

Proof.

$$\begin{aligned}
 \dots &= \operatorname{argmin}_{\pi} \mathbb{E}_{x \sim D} \left[\mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi^*(y|x)} \right] - \log Z(x) \right] \\
 &= \operatorname{argmin}_{\pi} \mathbb{E}_{x \sim D} [\mathbb{D}_{\text{KL}}(\pi(y|x) || \pi^*(y|x)) - \log Z(x)] \\
 &= \operatorname{argmin}_{\pi} \mathbb{E}_{x \sim D} [\mathbb{D}_{\text{KL}}(\pi(y|x) || \pi^*(y|x))] \\
 &= \pi^*(y|x)
 \end{aligned}$$

□

From (3), we get:

$$r(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \quad (4)$$

Apply (4) to (1), and $Z(x)$ cancels:

$$\therefore \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (5)$$

→ No longer a RL problem. No need of reward model. **Direct optimization**

+ Enhanced Stability

Future Research

So, is RL useless for finetuning language models?

We want to keep on communicating with LMs, tons of problems occur in **multi-turn** conversation:

ex) Snowballing Hallucination, Repeating, Forgetting, ...

+Lifelong learning language models, personalized to user

→Need a design for long-term goal