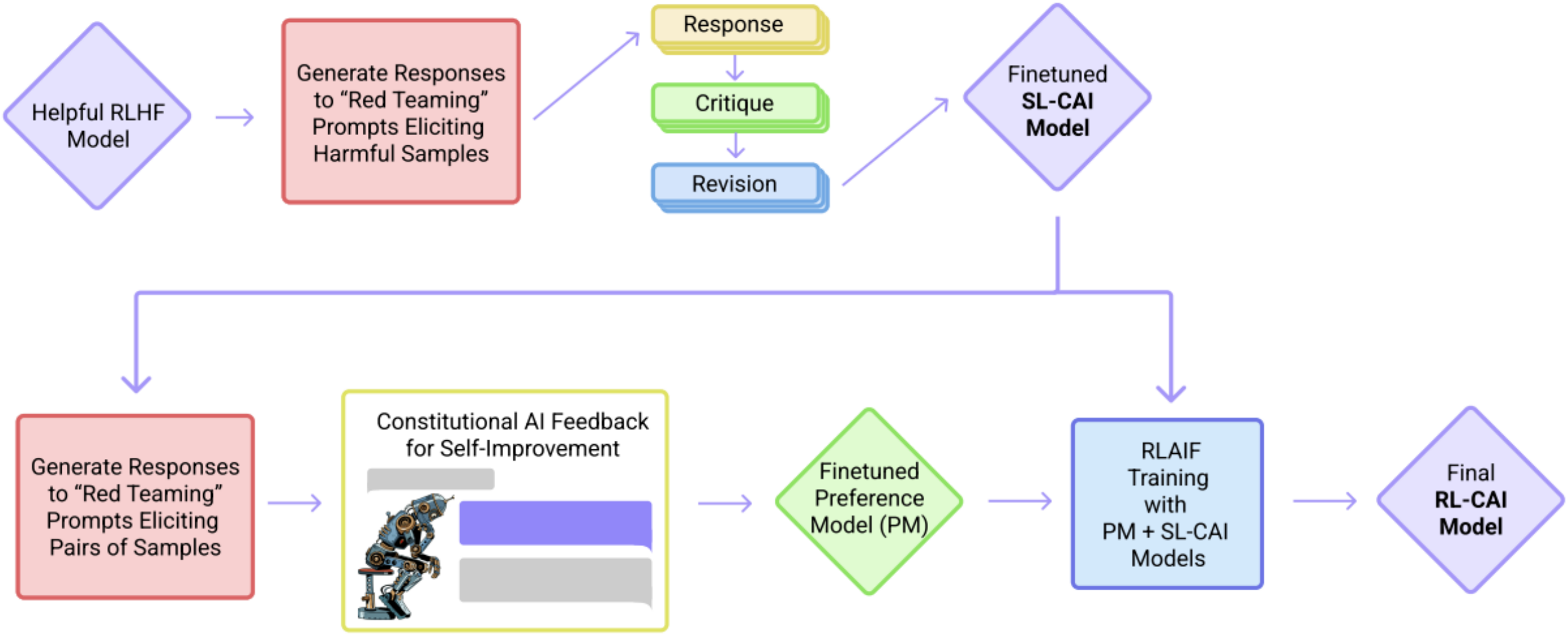
More Advanced RLHF

Discussion

Constitutional AI: Motivations

- AI supervision of AI systems (Scalable Oversight)
- Balancing harmlessness and helpfulness
- Make AI values/principles more transparent
- Reduce time and amount of human data when altering objective for an AI system

Constitutional AI



Example of revisions

- Section 3.1

Constitution

- Appendix C

Techniques used

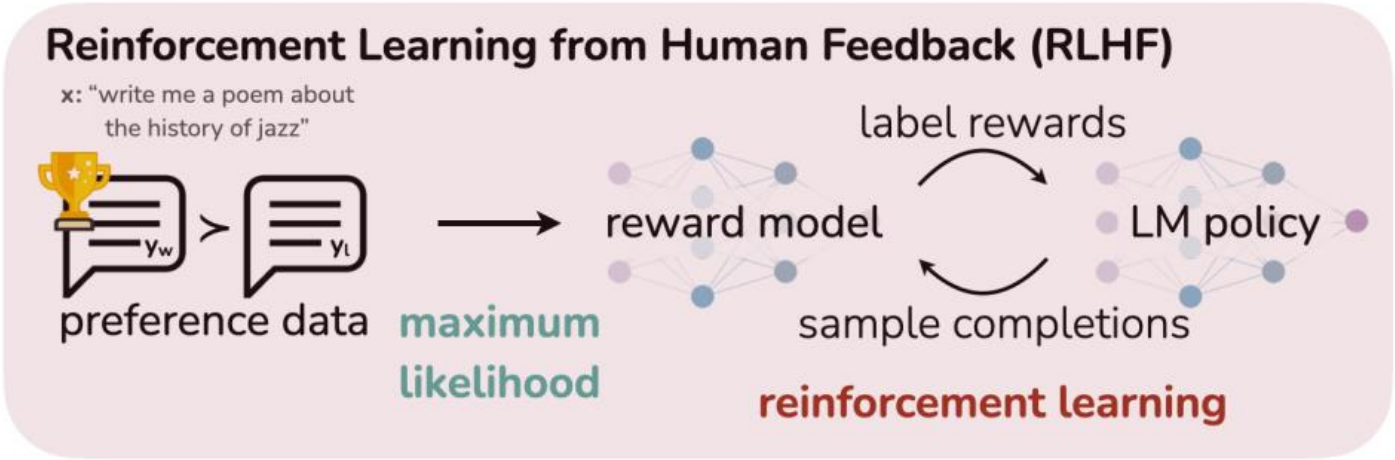
- Ensembling
- CoT
- Few shot prompting

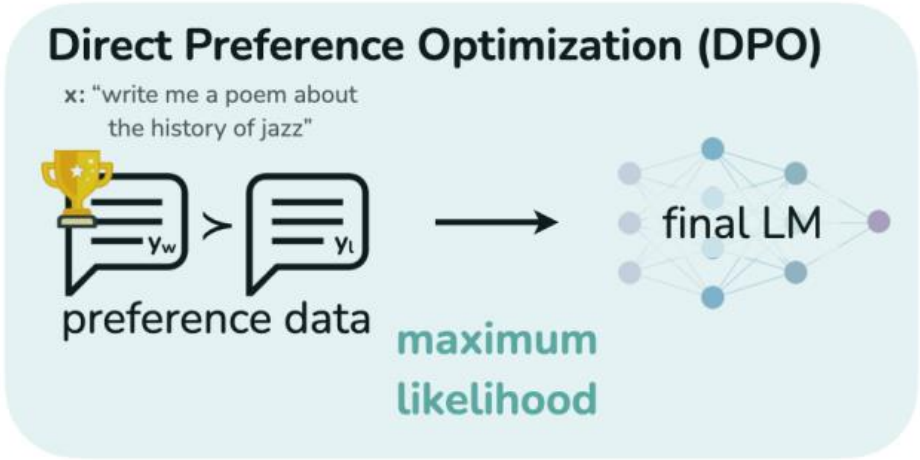
Discussion

- Pros and Cons?

- When is human data vs. AI data useful/harmful?

Direct Preference Optimization





Closed form of KL-constrained policy

- Objective:

$$\max_{\pi} \quad \mathbb{E}_{y\sim\pi(\cdot|x)}\left[r(x,y)\right] - \beta\, D_{KL}(\pi(\cdot|x) \,\|\, \pi_{ref}(\cdot|x))$$

Bigger beta means stay closer to reference policy

Optimal solution to above has this form (solve Lagrangian):

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

Small beta -> aggressive reward optimization
Large beta -> conservative, stay near reference

Deriving DPO objective

- Solve for reward instead of policy:

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

Deriving DPO objective

- Solve for reward instead of policy:

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

Take logs

$$\log \pi^*(y|x) = \log \pi_{ref}(y|x) + \frac{1}{\beta}r(x,y) - \log Z(x)$$

Rearrange to get r(x,y) by itself

$$r(x,y) = \beta \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x)$$

Deriving DPO

We still have this pesky partition function…

$$r(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x)$$

- Plug into Bradley-Terry Preference model

$$p(y_w \succ y_l|x) = \sigma\left(r(x, y_w) - r(x, y_l)\right)$$

DPO Derivation

$$r(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x)$$

- First calculate difference:

$$r(x, y_w) - r(x, y_l) =$$

$$= \beta \left[\log \frac{\pi(y_w|x)}{\pi_{ref}(y_w|x)} - \log \frac{\pi(y_l|x)}{\pi_{ref}(y_l|x)} \right]$$

DPO – Maximum Likelihood Objective

$$\max_{\theta} \quad \mathbb{E}_{(x,y_w,y_l)} \left[\log p_\theta(y_w \succ y_l | x) \right]$$

where

$$p(y_w \succ y_l | x) = \sigma \left(\beta \log \frac{\pi(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi(y_l|x)}{\pi_{ref}(y_l|x)} \right)$$

We have an implicit reward (defined by the policy) that we fit to data

$$\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$$

Discussion

- When might we prefer DPO over RLHF or vice versa?

- What are advantages or disadvantages to learning an explicit reward model?