reward model: learned from human prefs

↳ (x, y_w, y_l)

↳ $L(\theta_{RM}) = -\log P(y_w \succ y_l | x)$

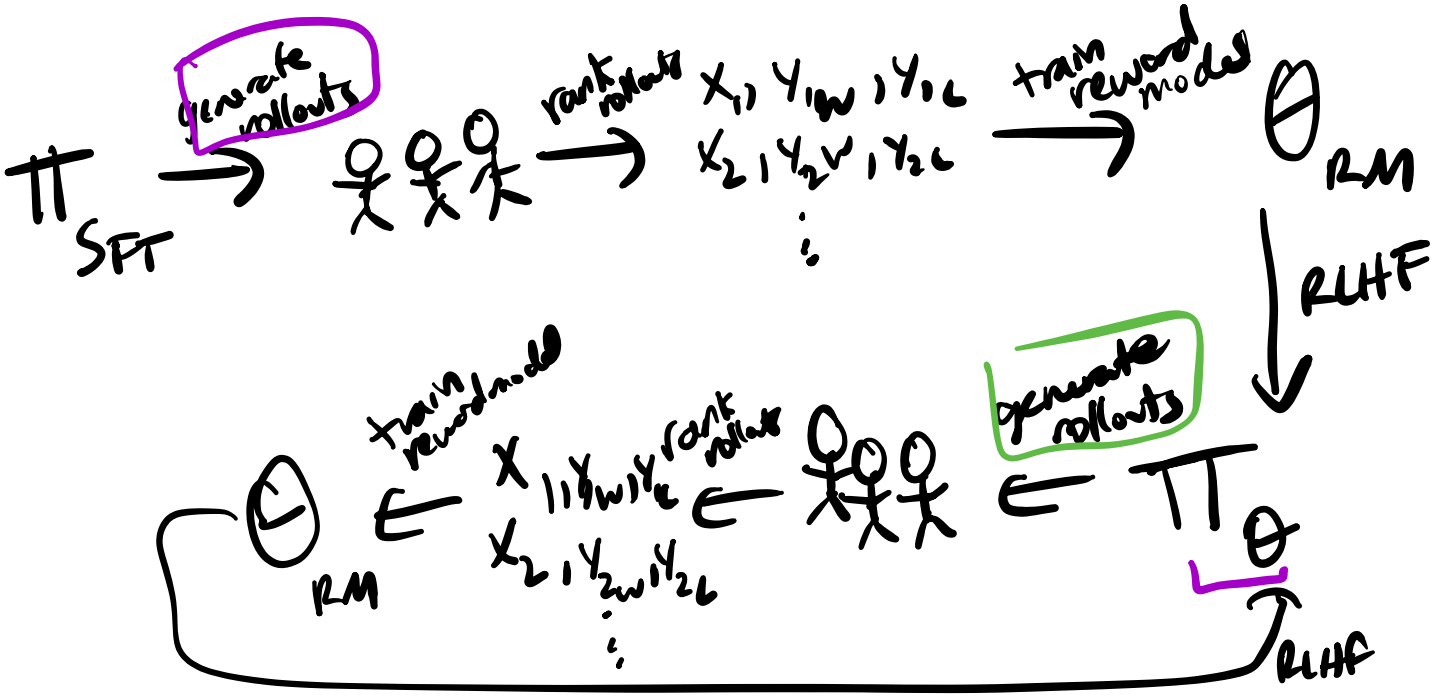
↳ $r_{RM}(x, y)$ plug into RL algo

PPO, GRPO, REINFORCE

reward hacking

↳ human prefs are noisy/subjective

↳ RM is only trained on rollouts from π_{SFT}



↳ how do we reduce the impact of reward hacking?

↳ prevent policy from drifting too far from the SFT checkpoint

↓

generalize to "reference" checkpoint

since we could start w/ π_{SFT}, π_{OPD}, $\pi_{\theta_{OLD}} \Rightarrow \pi_{ref}$

↓

checkpoint that produced rollouts for RM training

↳ use KL penalty to minimize drift

$$J(\theta) = \mathop{E}_{y \sim \pi_\theta(\cdot|x)}\left[r_{RM}(x,y)\right] - \beta\, KL\left(\pi_\theta(\cdot|x) \,\|\, \pi_{ref}(\cdot|x)\right)$$

↳ hyperparameter, controls weight of KL penalty

Comparing KL penalty to OPD:

↳ OPD student = π_θ, OPD teacher = π_{ref}

$$KL(\pi_\theta \| \pi_{ref}) = \mathop{E}_{y \sim \pi_\theta(\cdot|x)} \left[\sum_t \left(\log \pi_\theta(y_t | x, y_{<t}) - \log \pi_{ref}(y_t, x, y_{<t}) \right) \right]$$

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[-A_i \log \pi_\theta(y_i | x) + \beta \sum_t KL_t \right]$$

↳ if β too small ⇒ more reward hacking

↳ if β too large ⇒ π_θ stays very similar to π_{ref}

InstructGPT (Ouyang et al. 2022, OpenAI)

↳ started from GPT-3, 175B param model

↓

SFT, used 13K prompts w/ human-written responses ⇒ π_{SFT}

↓

trained reward model, sampling from π_{SFT}, 33K prompts

human labelers rank responses

↓

PPO, same set 33K prompts

reward : $r_{RM}(x,y) - \beta KL$

---

reward hacking / overoptimization

↳ length: RMs like longer answers, so if you swap out $r_{RM}(x,y)$ for length(y) as reward, you recover most gains

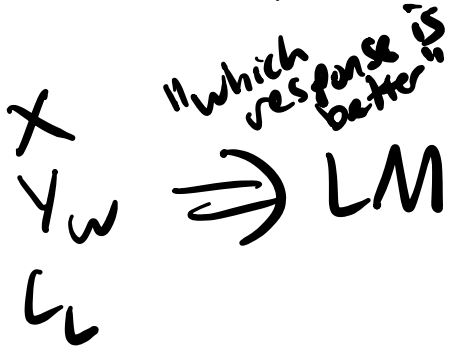
↳ sycophancy: humans prefer responses that they agree w/, so agreeing gets rewarded

↳ annotator distribution

RLAIF: RL w/ AI feedback

↳ only diff between RLAIF vs. RLHF is source of pref judgments

↳ instead of human labelers



↳ Anthropic "Constitutional AI"

↳ another approach: skip reward model entirely!

↳ reward = LLM("rate this output from 1-10")

↓

expand to structured rubrics