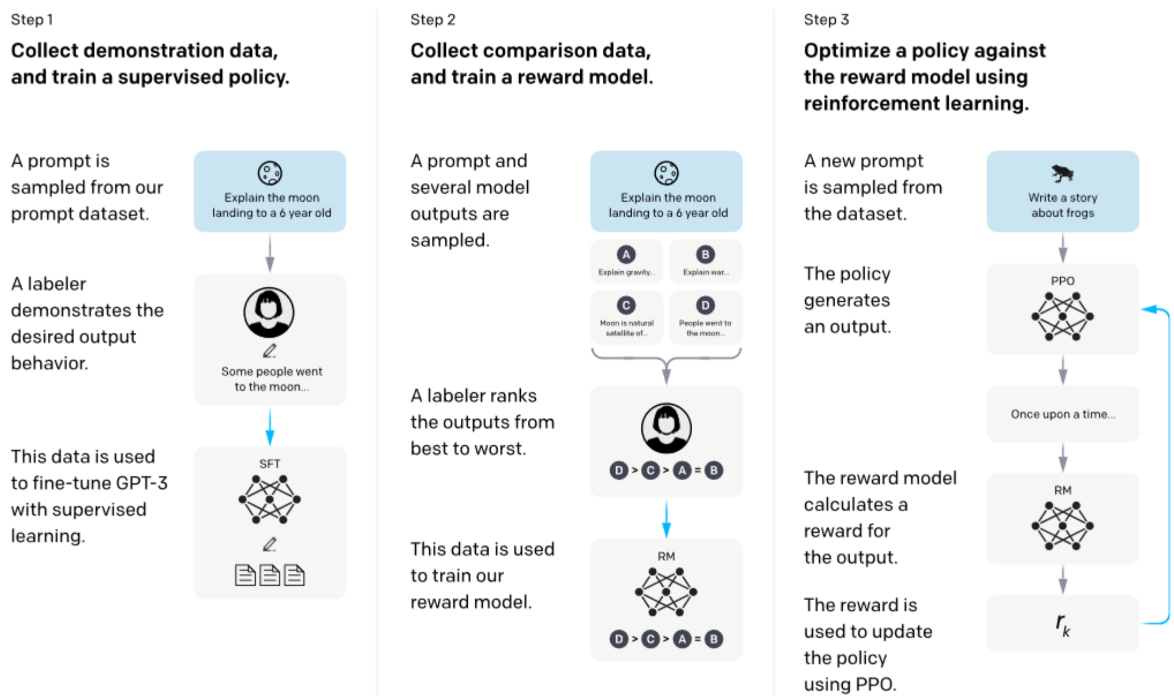
ChatGPT & InstructGPT

Training language models to follow instructions with human feedback (2022) 187

Goal: align language models with user intent

1. collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning
2. collect a dataset of **rankings** of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback

Result: improvements in truthfulness and reductions in toxic output generation



Step 1: human data → supervised fine-tuning (SFT)

Step 2: model outputs → reward modeling (RM)

Step 3: model after SFT + RM rewards → reinforcement learning (RL)

Reward modeling (RM): ranking → score

pairwise ranking loss for K responses

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

where $r_\theta(x,y)$ is the scalar output of the reward model for prompt x and completion y with parameters θ, y_w is the preferred completion out of the pair of y_w and y_l, and D is the dataset of human comparisons.

Reinforcement learning (RL): Proximal Policy Optimization (PPO)

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

where π_ϕ^{RL} is the learned RL policy, π^{SFT} is the supervised trained model, and D_{pretrain} is the pretraining distribution.

- $r_\theta(x,y)$: expected reward for the new model
- $\log(\pi_\phi^{\text{RL}}(y \mid x)/\pi^{\text{SFT}}(y \mid x))$: KL divergence to avoid going too far away from the original model
- $E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\text{RL}}(x))\right]$: objective for GPT3 on the original data