논문 발표

InstructGPT

김 용 진

INDEX

01

Abstract

Task

01. Bigger LM does not make them better at following a user' s intent

02. InstructGPT aligning language models with user intent

03. Fine–tuning with human feedback

04. Outputs from the 1.3B InstructGPT model are preferred to outputs from the 175B GPT–3

02

Introduction

LMs Defects

01. LMs often express unintended behaviors(biased, toxic, not following user instructions)

02. Because LLM predict the next token on a webpage from the internet

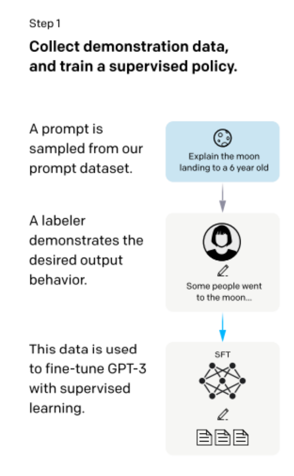
03. This is different from the objective "follow the user's instructions helpfully and safely"

Reinforcement Learning from Human Feedback Overview

01. Hire a team of 40 contractors to label data, based on their performance on a screening test

02. Collect a dataset of human–written demonstration of the desired output behavior on prompts

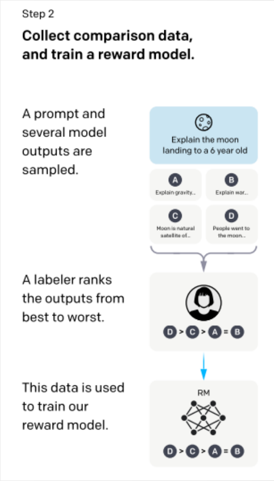
03. Use this to train supervised learning baselines



Reinforcement Learning from Human Feedback Overview

01. Collect a dataset of human-labeled comparisons

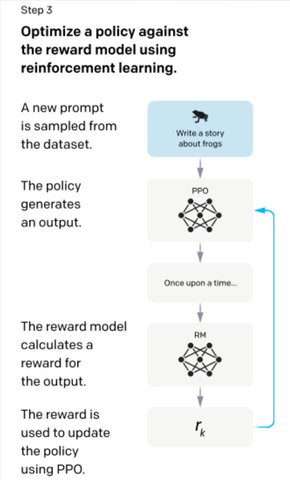
02. Train a reward Reward Model



Reinforcement Learning from Human Feedback Overview

01. Use RM as a reward function

01. Fine–tune Supervised learning baseline to maximize reward



Main Finding

01. Labelers significantly prefer InstructGPT outputs over outputs from GPT-3

02. InstructGPT models show improvements in truthfulness over GPT-3

03. InstructGPT shows small improvements in toxicity over GPT-3, but not bias

04. Minimize performance regression on public NLP dataset by modifying RLHF fine-tuning procedure

05. InstructGPT models show promising generalization to instructions outside of the RLHF fine-tuning distribution

06. InstructGPT shows small improvements in toxicity over GPT-3, but not bias

03

Method

Task

01. Training tasks are from two sources

02. Dataset of prompts written by labelers

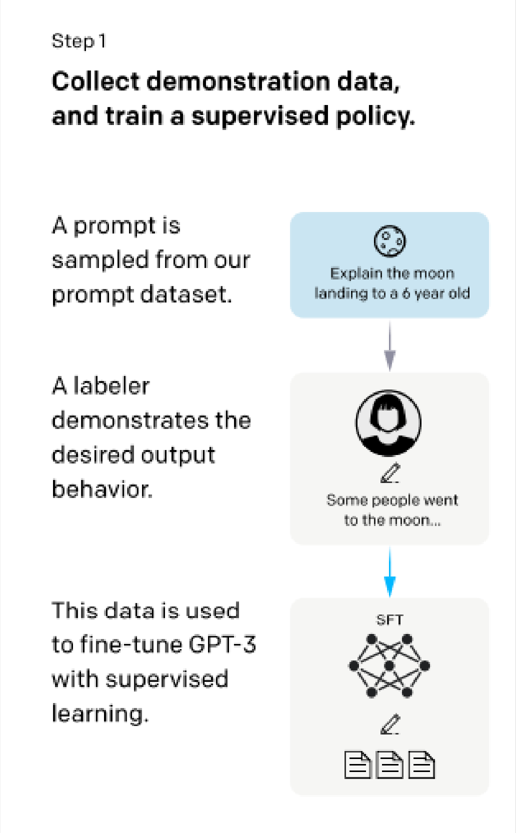
03. Early InstructGPT models

Supervised fine-tuning(SFT)

01. Fine–tune GPT–3

02. Final SFT model selection based on the RM score in Val set

03. Training more epochs help both RM score and human preference rating



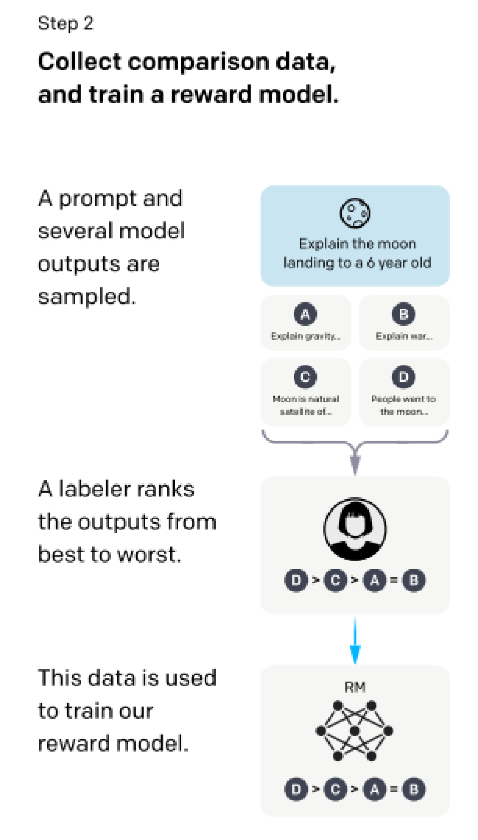
Reward Models(RM)

01. Starting from the SFT model with the final unembedding layer removed

02. Model take prompt and response, output a scalar reward

03. RM is trained on a dataset of comparisons between two model outputs on the same input

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

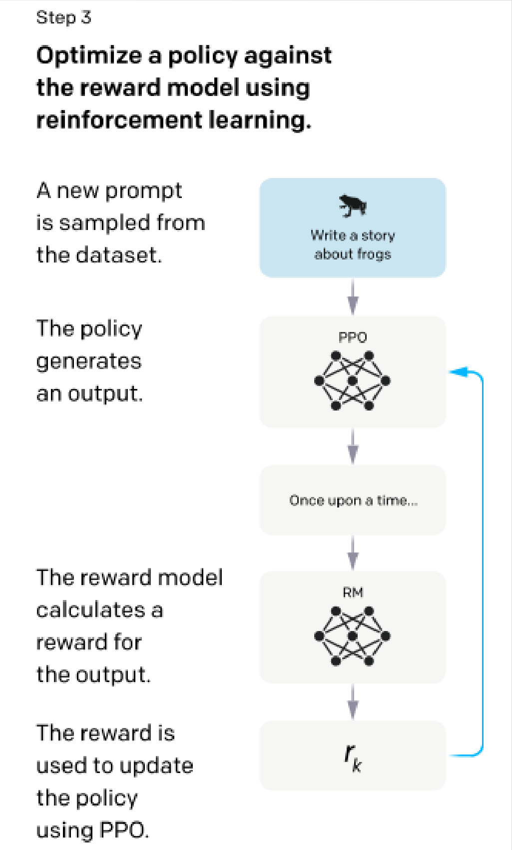


Reinforcement Learning

01. Fine-tuned the SFT model on environment using PPO

02. Random customer prompt and expects a response to the prompt

03. Given the prompt and response, it produces a reward

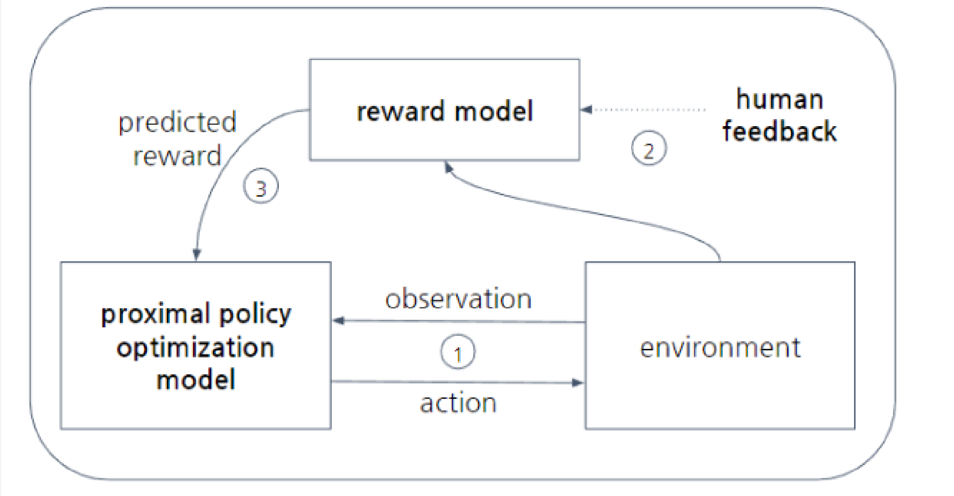


PPO: Policy based Model

Agent: PPO Model

Environment: Input Sentence

Action: Output Sentence



Do train if 0.8 ~ 1.2

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

Extract x from pretrain dataset

01. Need human labeling

02. No improvement in toxic, bias

THE END

감 사 합 니 다