Train LM to follow instructions with HF: InstructGPT

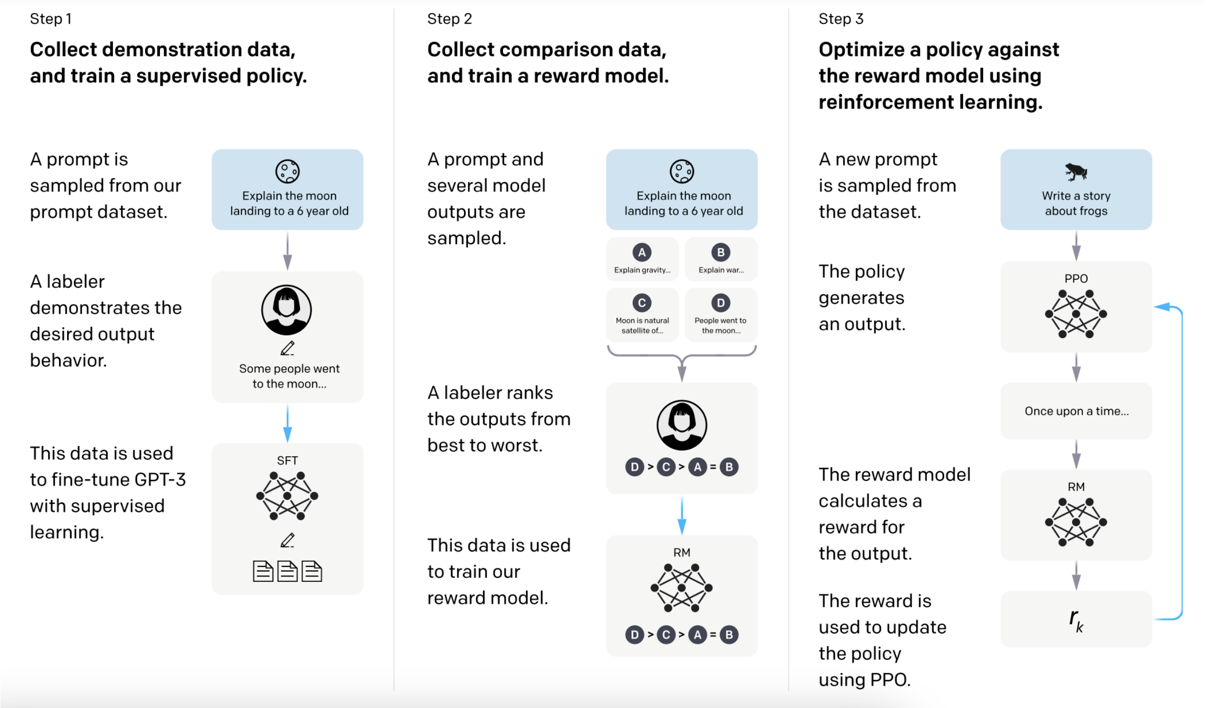
Misaligned objective: “predict next token on a webpage in internet” different from “follow user’s instruction helpfully and safely”

LMs should follow instructions, and be truthful.

Use RLHF with GPT 3 to follow class of written instructions, human preferences as a reward signal to fine-tune models,

InstructGPT is funetuned on human data.

- reward model (RM) on this human written desired outputs dataset to predict which model output our labelers would prefer,
- RM is used as a reward functions to train supervised baseline model,
- then researchers rate the quality of model outputs.


Step 1
Collect demonstration data, and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain the moon landing to a 6 year old
A labeler demonstrates the desired output behavior.
Some people went to the moon...
This data is used to fine-tune GPT-3 with supervised learning.
SFT
Step 2
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled.
Explain the moon landing to a 6 year old
A Explain gravity...
B Explain war...
C Moon is natural satellite of...
D People went to the moon...
A labeler ranks the outputs from best to worst.
D > C > A = B
This data is used to train our reward model.
RM
D > C > A = B
Step 3
Optimize a policy against the reward model using reinforcement learning.
A new prompt is sampled from the dataset.
Write a story about frogs
The policy generates an output.
PPO
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k

1: Supervised Finetuning, 2: Reward Function, 3: Proximal Policy Optimization

- Labellers prefer few shot instructions' InstructGPT 1.3B and 175B over normal GPT3
- Lesser halucinations, (a 21% vs. 41% hallucination rate) means more truthfullness.
- It is less toxic when prompted to be respectful,
- there is performance regressions on public NLP datasets, but it can be minimised, by mixing PPO updates with updates that increase the log likelihood of the pretraining distribution (PPO-ptx)

> ℹ️ PPO updates refer to the application of the PPO algorithm to fine-tune the model, and Log likelihood is a measure of how likely the model assigns a specific token to a given context, **The proposed approach, PPO-ptx, combines PPO updates with updates that aim to increase the log likelihood of tokens according to the pretraining distribution.**

- The model summarizes code, answers code-related questions, and can understand instructions in different languages. GPT3, on the other hand, struggles with these tasks and needs to be carefully prompted.

Related Work: Studies show that fine-tuning LMs on various NLP tasks, with instructions, enhances their performance on held-out tasks, in both zero-shot and few-shot settings.

LMs are toxic, have bias, but interventions in training dataset can reduce ability to model text from under-represented groups.

Train LMs with

- small, value-targeted dataset,
- removing documents in dataset, which has a high conditional likelihood for generating words,
- data filtering, blocking certain words or n-grams during generation, safety-specific control tokens

Experimentations

1. Use distribution of prompts which produce aligned outputs, done by labelers, and train a supervised model,

2. labels rank model outputs they prefer for an input, train a reward model on this dataset,
3. Fine-tune the supervised policy to optimize this reward using the PPO.

2, and 3 can be done multiple times to get better datasets and policy.

In the dataset**(they are specific and also have examples)** the authors deduplicated prompts by checking for prompts that share a long common prefix, and decided on three prompts

- plain prompt for arbitrary task with diversity,
- few shot,
- matching some usecases, received from OpenAI waitinglist applications.

Models

- SFT- Supervised Fine Tuning, overfitted, but does well on RM task and human preferences,
- RM model, take prompt and answer and output reward, labellers were asked to rank 4-9 responses, basically K^2 combinations, and this was used as an single example to train the model.
- RL, PPO in an environment, where a prompt is given, and response is expected, and RM model is used to generate rewards, there is also a per token penalty from SFT model to mitigate over optimization of RM.

InstructGPT is compared to 175B GPT-3 on the FLAN, trained on various NLP tasks and has NL instructions.