

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is
sampled from our
prompt dataset.



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.



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.



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