Learning from human feedback is crucial for modern ML systems!


2

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.

This data is used to fine-tune GPT-3.5 with supervised learning.

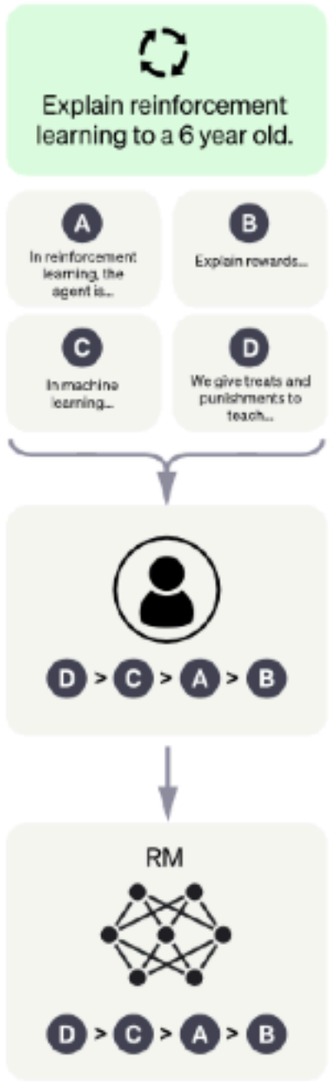
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

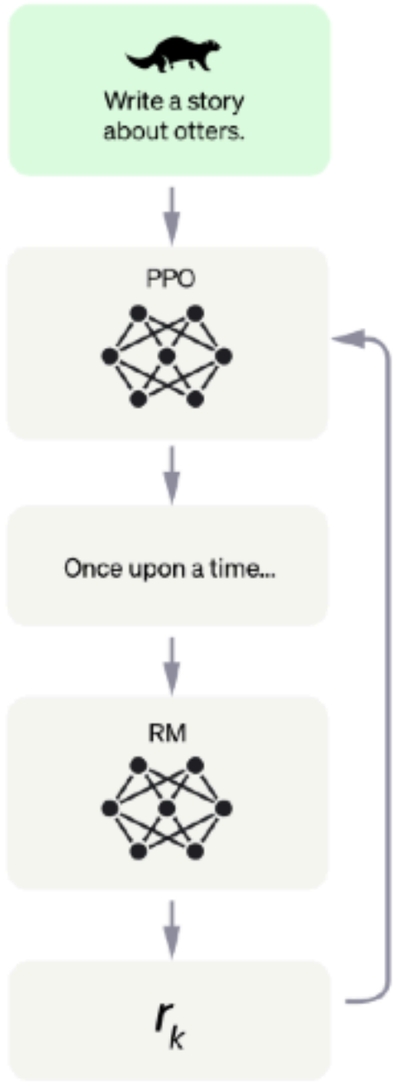
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Which response do you prefer?

Your choice will help make ChatGPT better.

Response 1

Absolutely! Here are some tips to make your brainstorming sessions more effective:

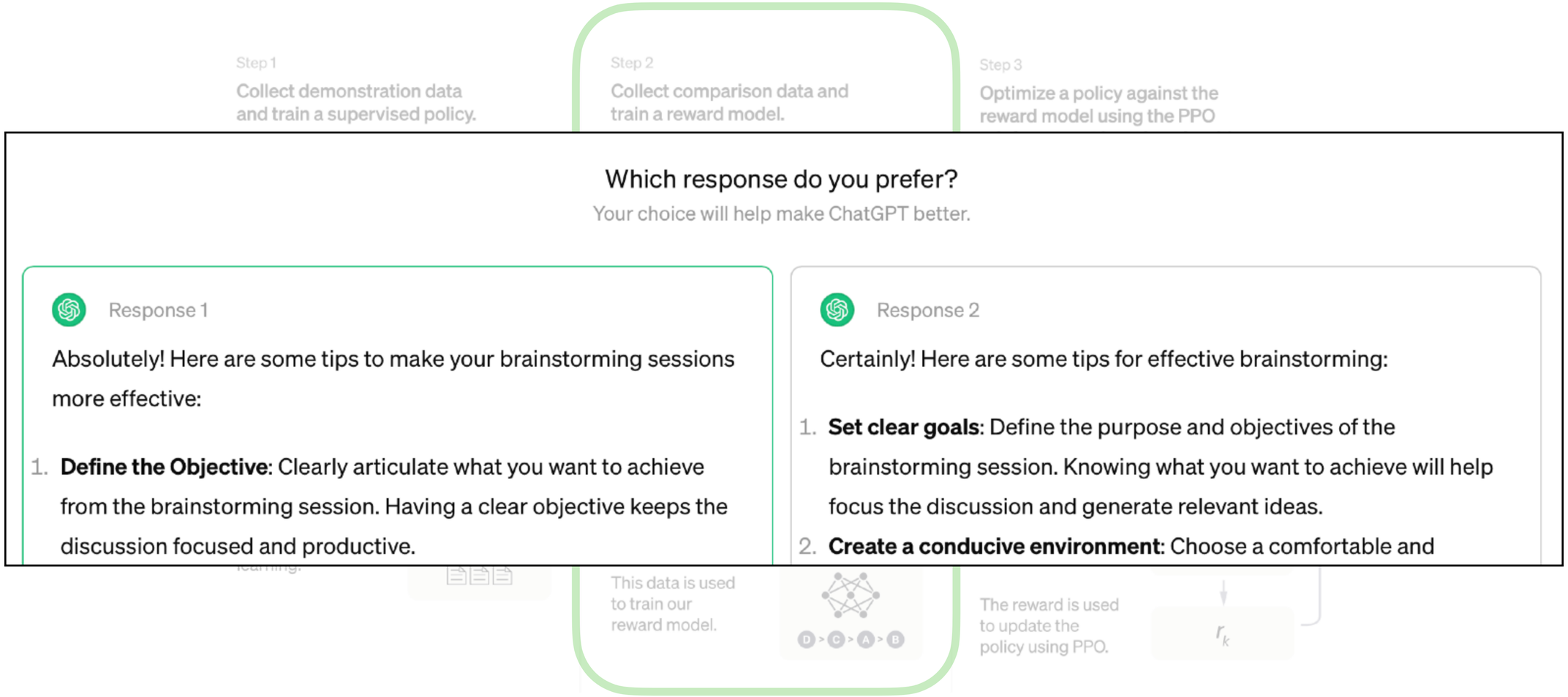
1. **Define the Objective**: Clearly articulate what you want to achieve from the brainstorming session. Having a clear objective keeps the discussion focused and productive.

Response 2

Certainly! Here are some tips for effective brainstorming:

1. **Set clear goals**: Define the purpose and objectives of the brainstorming session. Knowing what you want to achieve will help focus the discussion and generate relevant ideas.
2. **Create a conducive environment**: Choose a comfortable and

Learning from human feedback is crucial for modern ML systems!



How should machines learn from humans?

- Volume of responses needed is **massive!** → time, money, infrastructure

*How can we more effectively learn **with** human feedback and when can we use foundational tools to learn **without** additional feedback*