Natural Language Processing

ICS 491

We all know about ChatGPT

How does it work?

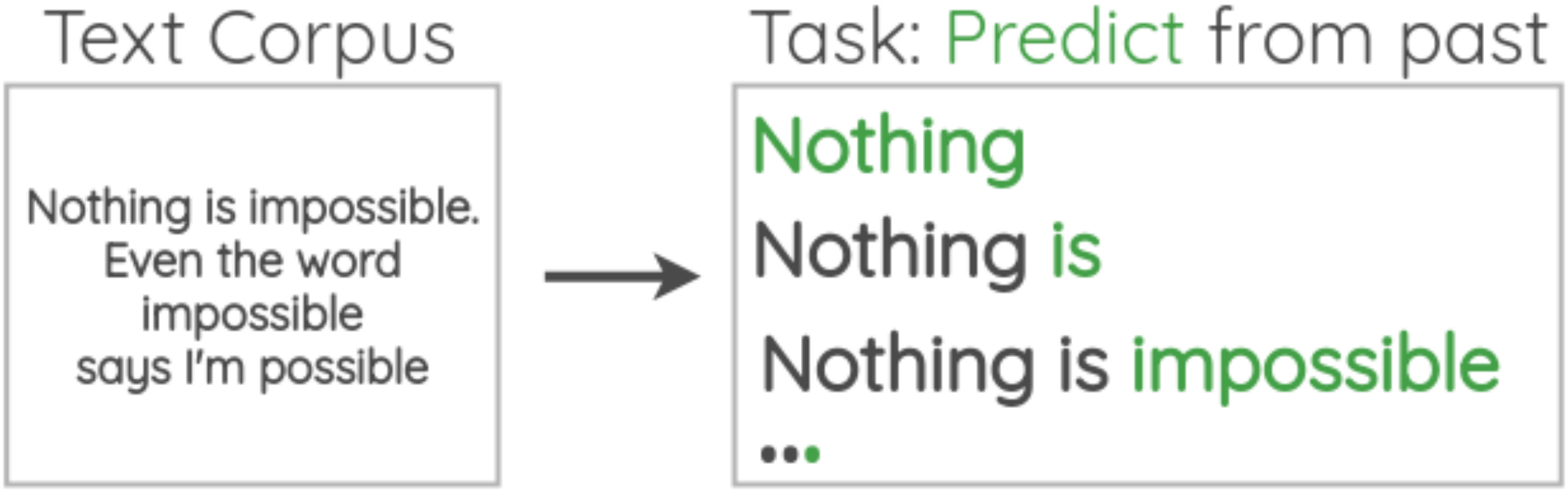
Core Ingredients of ChatGPT

1. Self-supervised learning on the whole Internet

2. Fine-tune model using Reinforcement Learning from Human Feedback (RLHF)

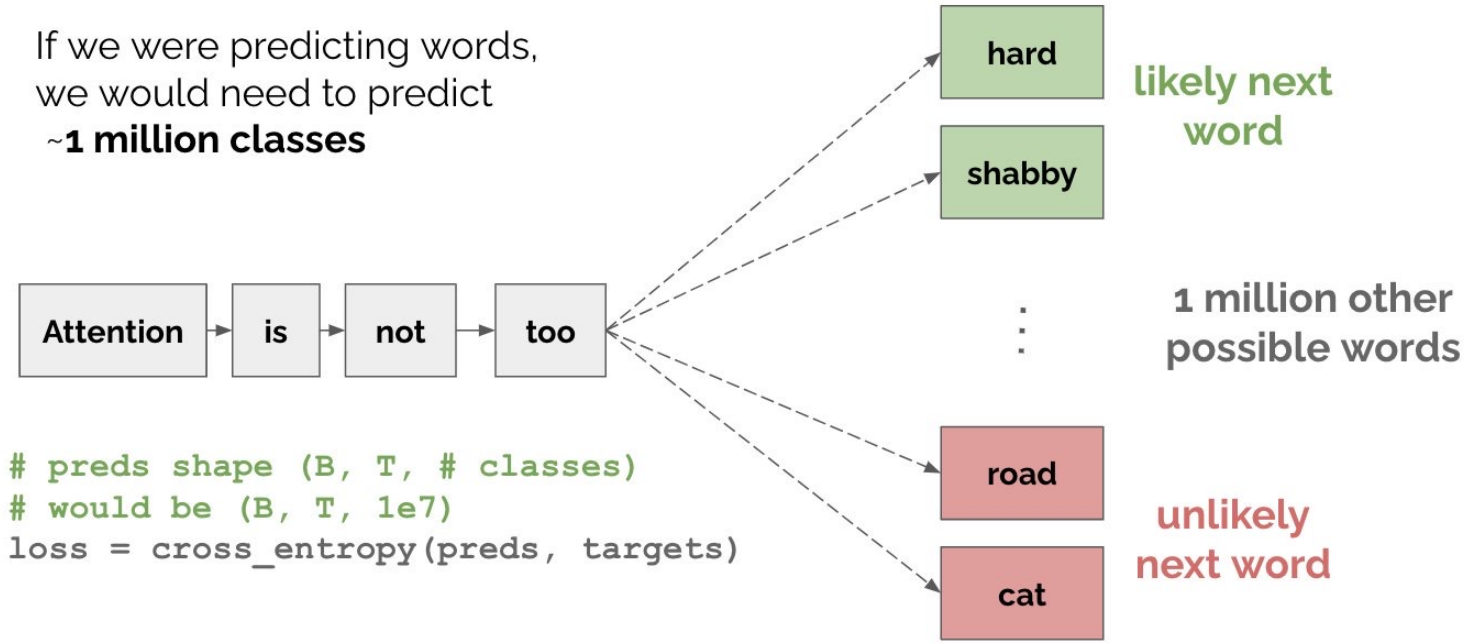
What is Self-Supervised Learning?

- Simple idea: Train a model to make predictions of labels that are not provided by humans, like the next word in a sentence (no human labels required)
- Enables training on all the text possible



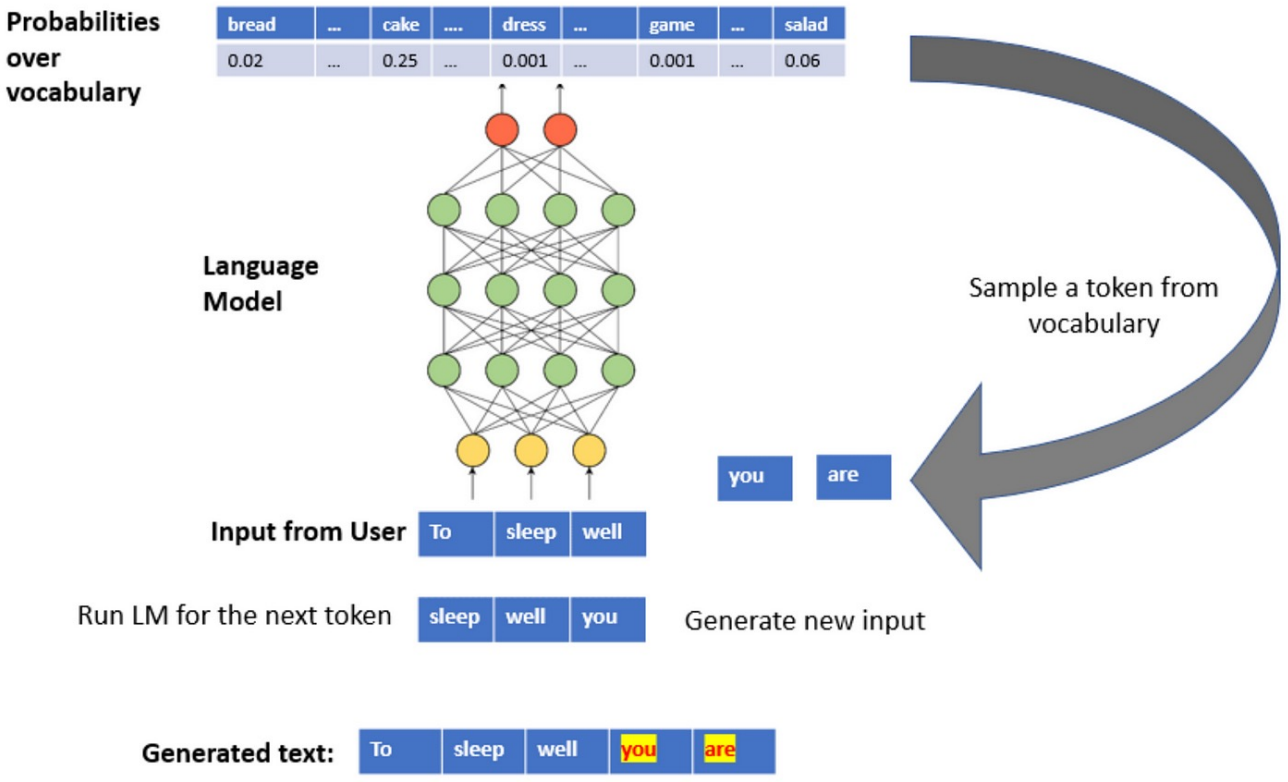
The power of Self-Supervised Learning

- If you have a model that can predict the next word given previous words, then you have a model that understands human language pretty well

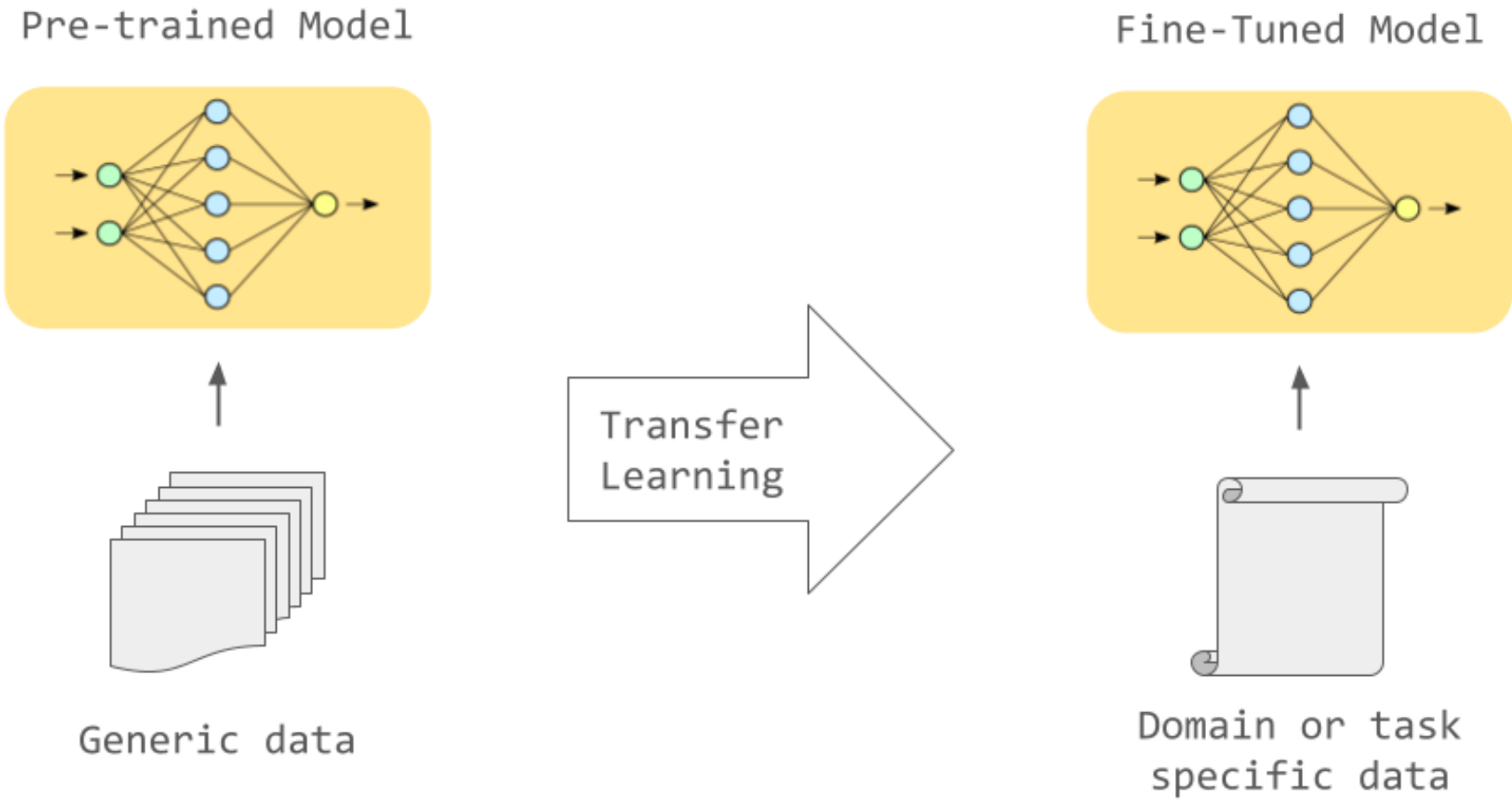


The power of Self-Supervised Learning

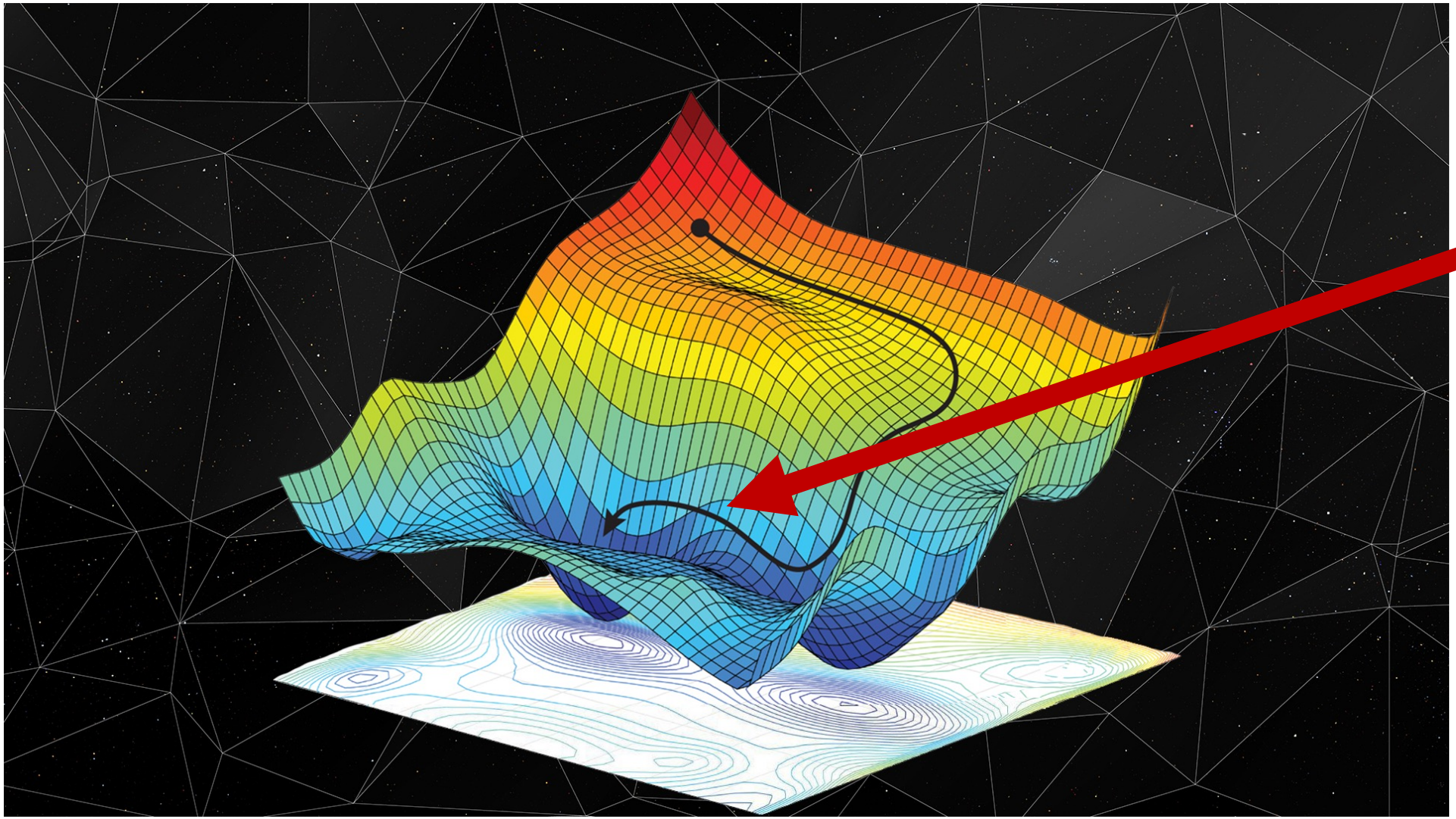
- If you have a model that can predict the next word given previous words, then you have a model that understands human language pretty well



Next step: fine-tuning



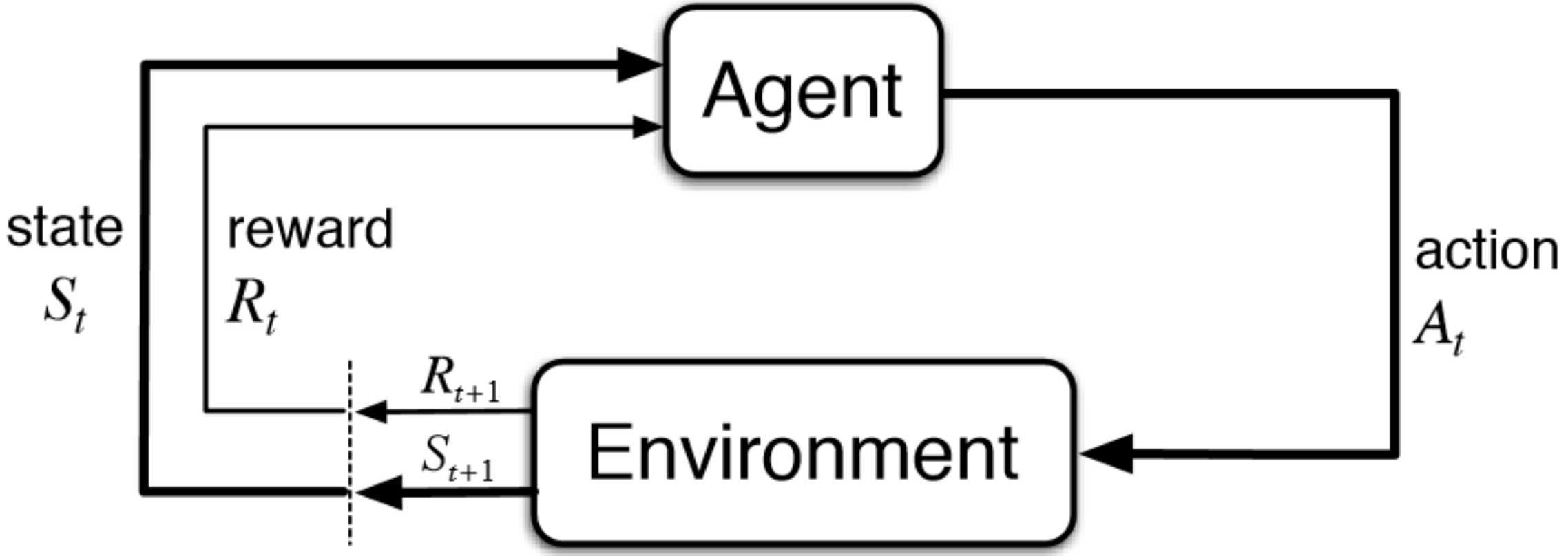
Next step: fine-tuning



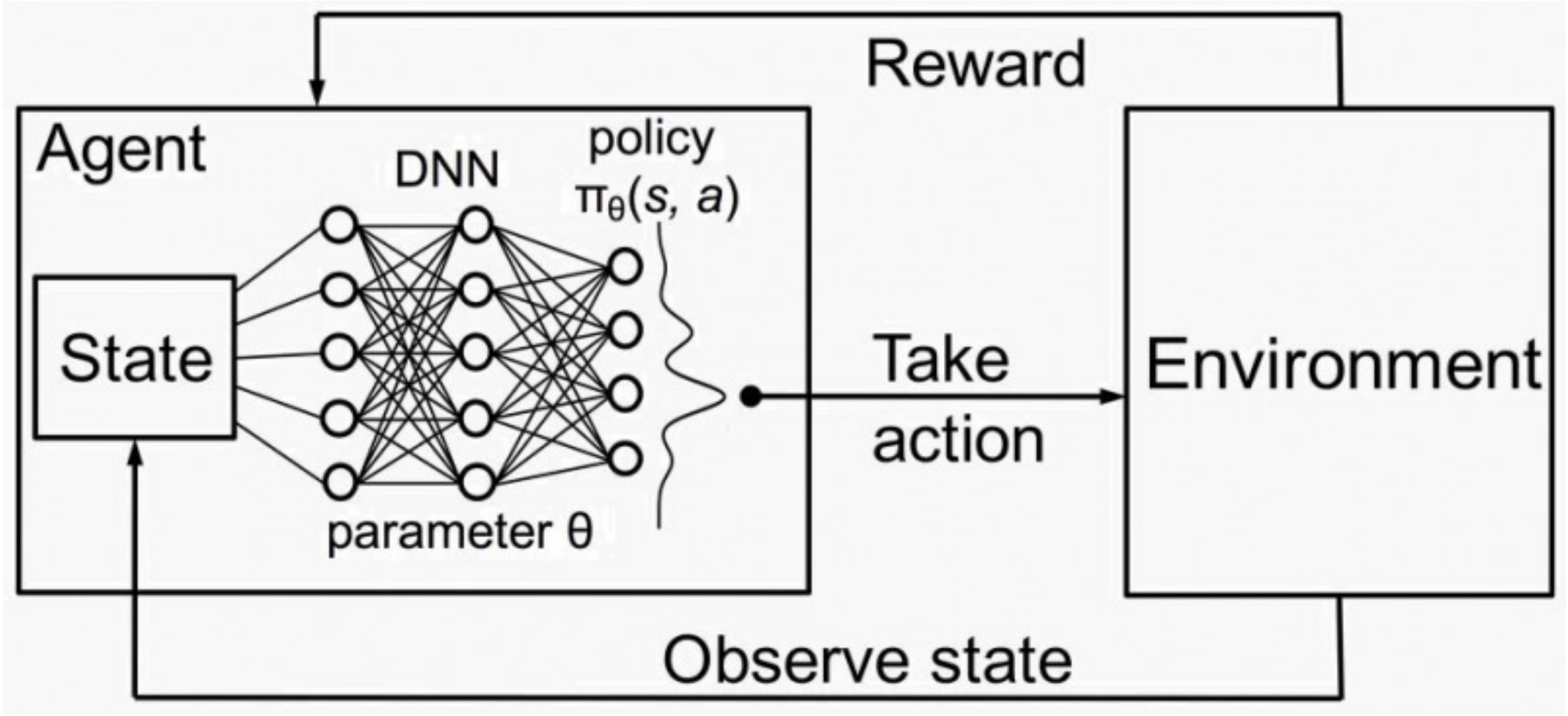
We start much closer to where we need to end up when we pre-train using self supervised learning

Fine-tuning with Reinforcement Learning

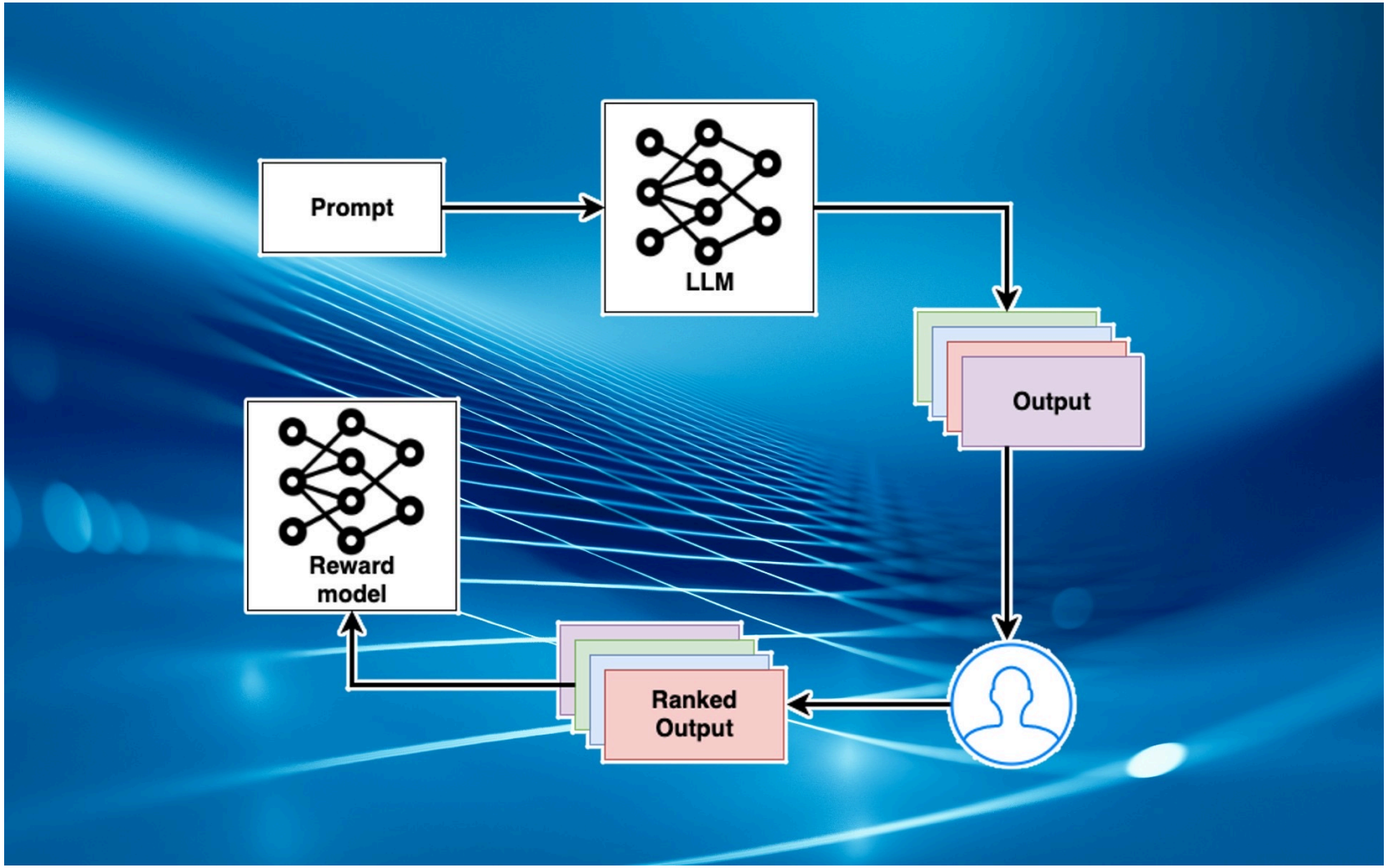
What is Reinforcement Learning at a very high level?



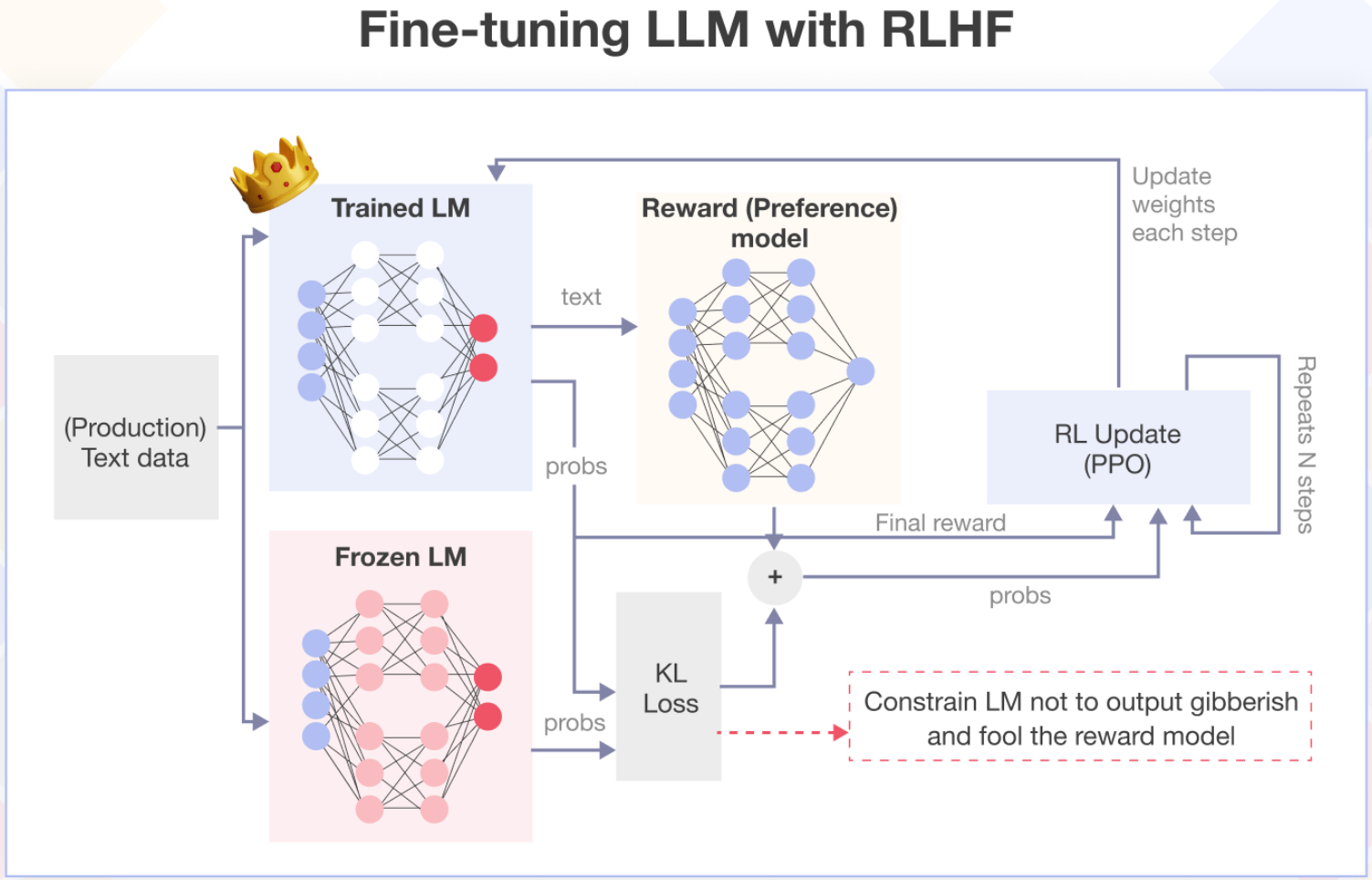
ChatGPT uses **Deep** Reinforcement Learning



Reinforcement Learning with Human Feedback (RLHF)



Fine-tuning with RLHF



Fine-tuning with RLHF

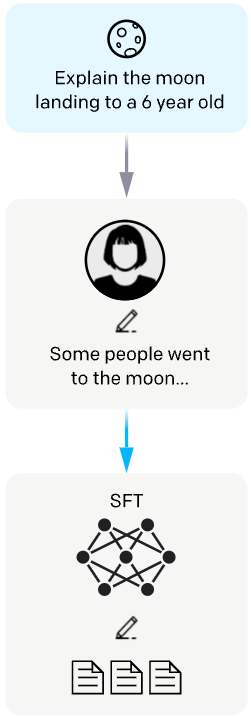
Step 1

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 with supervised learning.



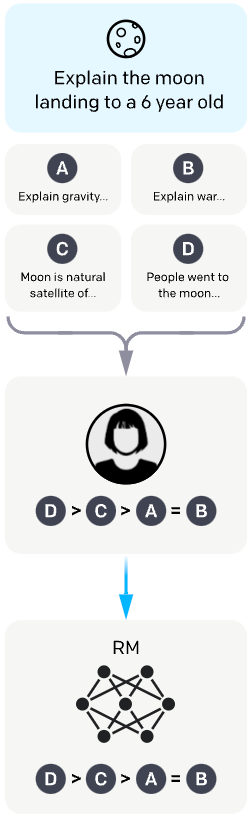
Step 2

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

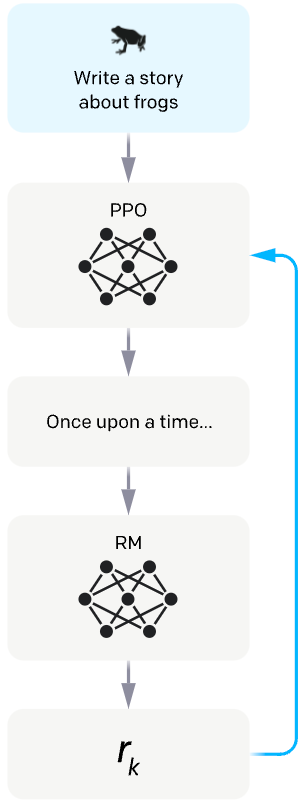
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Core Ingredients of ChatGPT

1. Self-supervised learning on the whole Internet

2. Fine-tune model using Reinforcement Learning from Human Feedback (RLHF)

This paradigm can probably be applied to many future applications…

This paradigm can probably be applied to many future applications…!

Reminder about next few weeks

Tue Nov 21	Social Network Analysis (Class on Zoom)	
Thu Nov 23	Thanksgiving Holiday (No Class)	
Tue Nov 28	Watch final project videos (No Class)	Project Milestone #6: Final Presentation
Thu Nov 30	Watch final project videos (No Class)	
Tue Dec 5	Multimedia Analytics	
Thu Dec 7	Course Overview	
Fri Dec 15		Final Project Infographic and Code