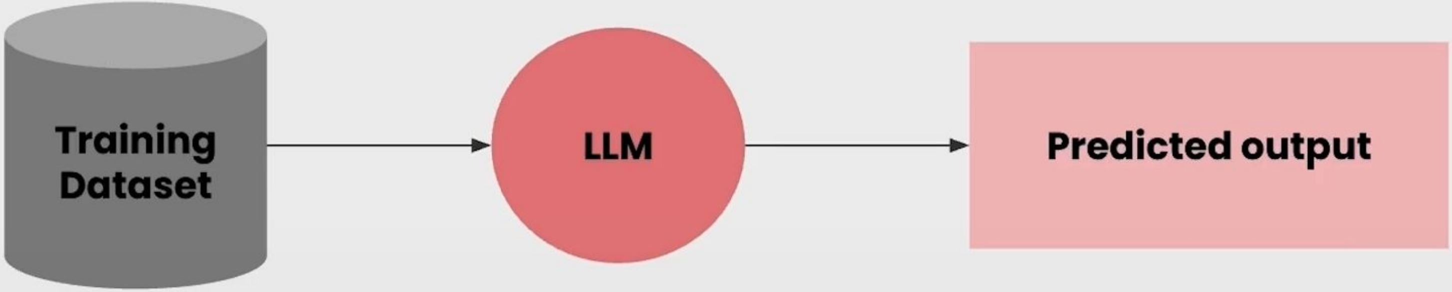
Predibase

DeepLearning.AI

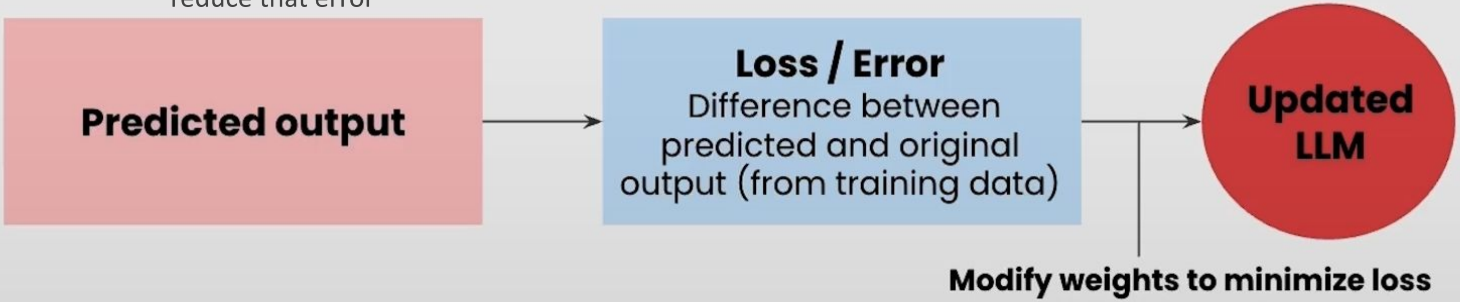
Supervised Fine-Tuning (SFT)

Learning From Past Demonstrations

Forward Pass The model generates an output for the given prompt



Backward Pass Compute the error by comparing the model's output to the correct response and update the model's weights to reduce that error



SFT Dataset

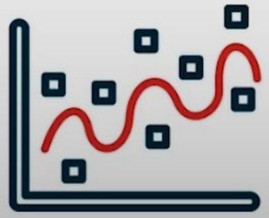
Prompt	Completion
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?	72
Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?	10
James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?	624

SFT Dataset + Reasoning

Prompt	Completion
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?	**<think>**Natalia sold 48/2 = <<48/2=24>>24 clips in May. Natalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May**</think>** **<answer>**72**</answer>**
Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?	**<think>**Weng earns 12/60 = $<<12/60=0.2>>0.2 per minute. Working 50 minutes, she earned 0.2 x 50 = $<<0.2*50=10>>10.**</think>** **<answer>**10**</answer>**
James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?	**<think>**He writes each friend 3*2=<<3*2=6>>6 pages a week So 6*2=<<6*2=12>>12 pages every week, or 12*52=<<12*52=624>>624 pages a year**</think>** **<answer>**624**</answer>**

Limitations of Supervised Fine-Tuning

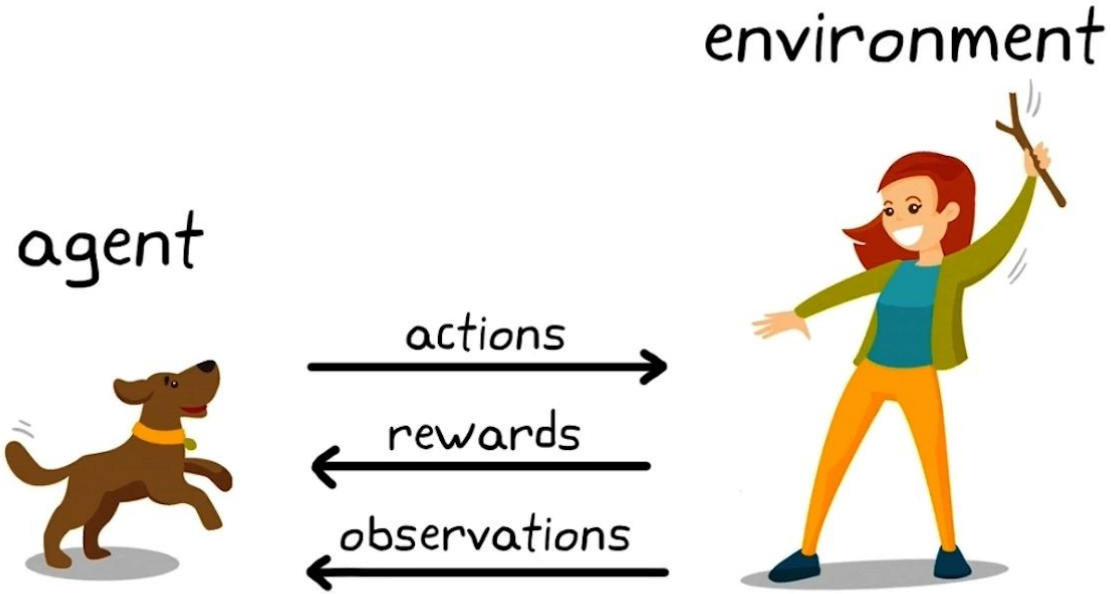
Heavily relies on large (1000s of examples), accurately labeled datasets.



Model may learn patterns too specific to the training domain (overfitting).

Reinforcement Learning: Intuition

Reinforcement Learning: An agent learns to complete a task by trying different actions and receiving rewards, with goal of getting the most reward over time.



Action: Puppy performs trick.

Reward: You give or withhold a treat.

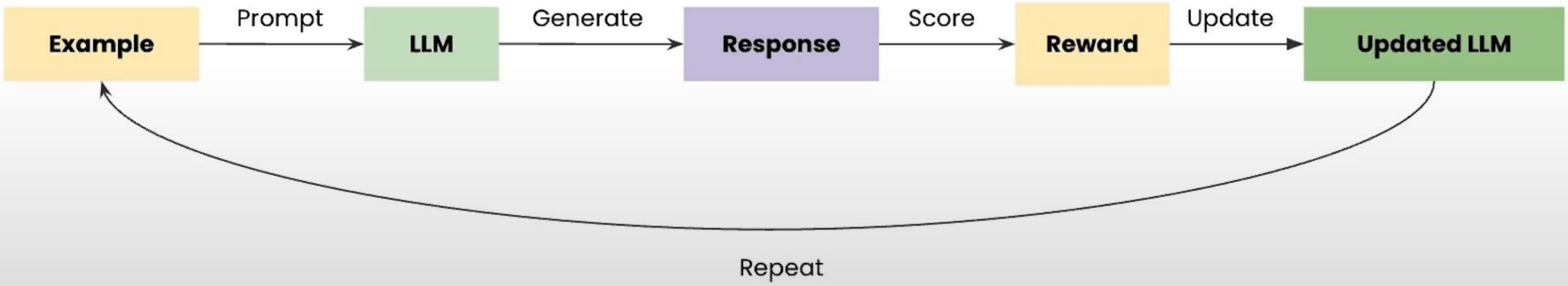
Observation: Puppy notices the result and updates its memory.

Image Source: Mathworks

Predibase

DeepLearning.AI

Applying Reinforcement Fine-Tuning To LLMs



Algorithms for Reinforcement Learning

1. Reinforcement Learning with Human Feedback (RLHF)
2. Direct Policy Optimization (DPO)
3. Group Relative Policy Optimization (GRPO)

Reinforcement Learning with Human Feedback (RLHF)

Step 1: Generate responses

Prompt

LLM

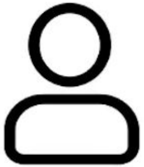
Response A

Response B

Response C

Response D

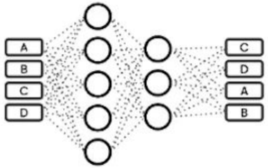
Step 2: Get human feedback



Human ranks responses:

C > D > A > B

Step 3: Train reward model



Step 3 details
Train a separate reward model to learn to predict these human preferences using a prompt and response pair as the input value & human ranks responses as target

Step 4 details
For each prompt, the LLM generates a response, the reward model scores it, and the LLM's weights are updated to increase the likelihood of producing high-scoring outputs. As this step is repeated over hundreds of prompts, it leans to generate responses that will produce high scores and align with human preferences.

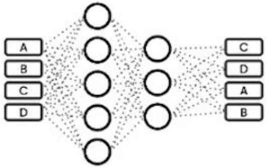
Step 4: Update LLM weights using reward model and PPO algorithm

Prompt

LLM

PPO updates weights

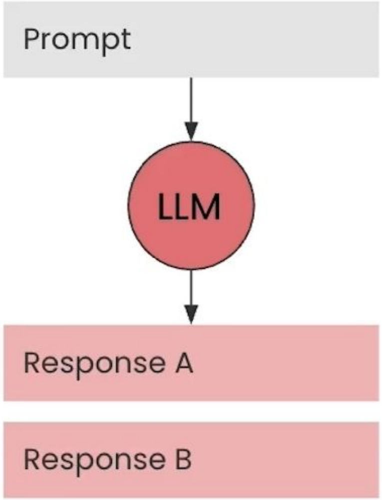
Response A



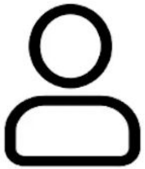
Reward score

Direct Policy Optimization (DPO)

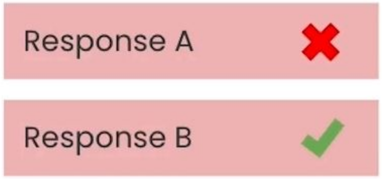
Step 1: Generate responses



Step 2: Get human feedback



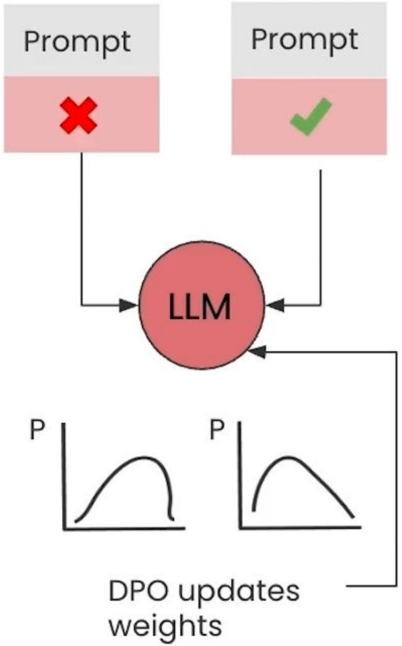
Human picks preferred response:



Step 3: Create preference dataset of many examples

Prompt	Chosen ✓	Rejected ✗

Step 4: Update LLM weights using preference dataset and DPO algorithm



Challenges With RLHF and DPO

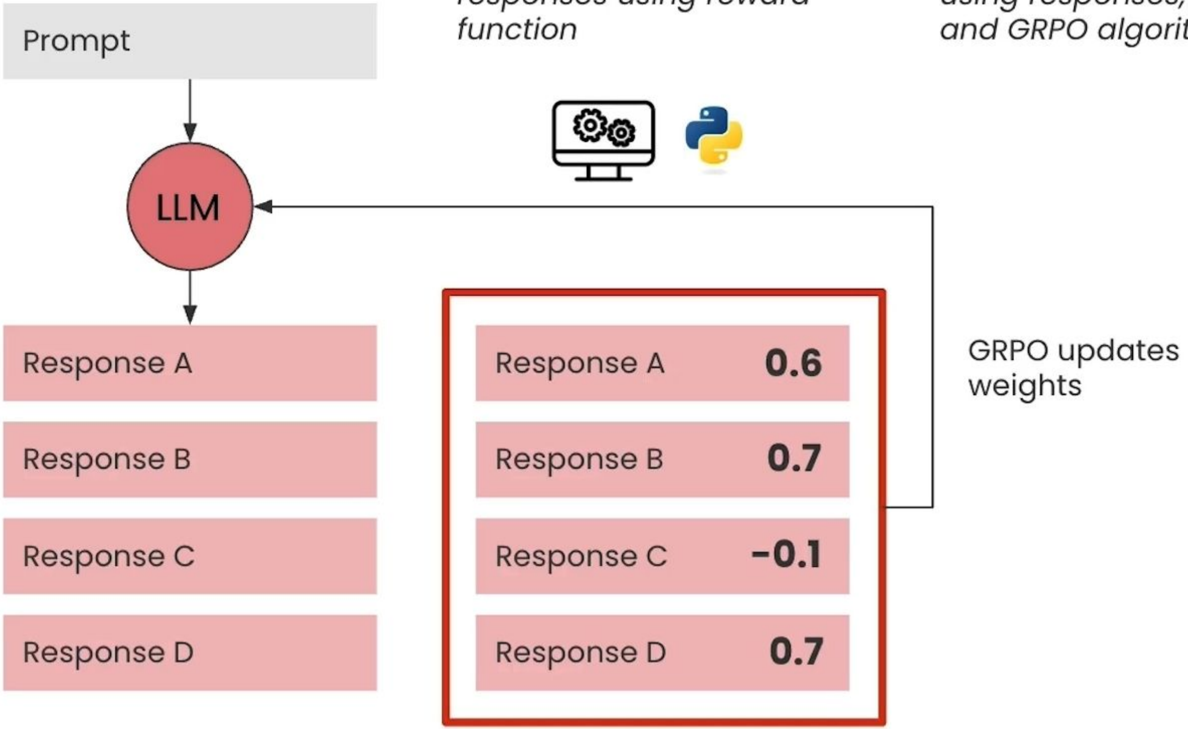
Challenge	RLHF	DPO
Data Needed	Ranked generations for reward model	Paired-preference labels (A > B)
Compute / Memory	Very high	Moderate
Training Stability	Often unstable—reward hacking, collapse risk	More stable, but still needs lots of labels
Limitation	Doesn’t teach new tasks—only steers preferences	Doesn’t teach new tasks—only steers preferences

Group Relative Policy Optimization (GRPO)

Step 1: Generate responses

Step 2: Assign scores to responses using reward function

Step 3: Update LLM weights using responses, scores, and GRPO algorithm



Many more details on reward functions and GRPO algorithm later in the course!

Benefits of Reinforcement Fine-Tuning

✓ Doesn't require labeled data, just a means to "verify" correctness

✓ Works with as few as 10 examples

✓ More flexible than SFT because it learns *actively* from feedback rather than from *fixed* labeled examples

✓ Enables reasoning models to organically discover better strategies by improving its chain of thought

When should you use Reinforcement Fine-tuning?

RFT can work well in situations where...

1. **You have no labeled data...**

...but you can *verify* the correctness of the output (e.g. transpiling source code).

2. **You have limited labeled data...**

...but not enough for supervised fine-tuning (i.e. generally less than 1000 labeled examples).

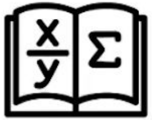
3. **Chain-of-thought reasoning improves performance**

Task performance improves significantly when you apply chain-of-thought reasoning.

Predibase

DeepLearning.AI

Tasks suited for Reinforcement Fine-tuning

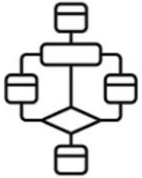


Mathematical Problem Solving

RF-tuned models can show their work, providing detailed reasoning behind calculations rather than just giving an answer.

Code Generation and Debugging

They understand the structure and intent of code, making them invaluable for AI-assisted programming.



Logical and Multi- Step Reasoning

Unlike simpler models that rely on pattern matching, reasoning models explain their decisions with step-by-step logic.

When should you choose RFT over SFT?