
DAY 1 · LANGUAGE MODELS · PART 5 OF 5

From GPT to ChatGPT

Instruction tuning, reward models and RLHF

Pavlos Protopapas

Harvard University



Outline

- › Transformers (Recap)
- › Instruction-Tuned LLMs (How Does ChatGPT Work?)
 - › Instruction Training Overview
 - › RLHF (Reinforcement Learning from Human Feedback)
 - › PPO (Proximal Policy Optimization)
 - › Clipping

Transformers

```
1 prompt = "Cheese is a fascinating food because "  
2 max_tokens = 1000
```

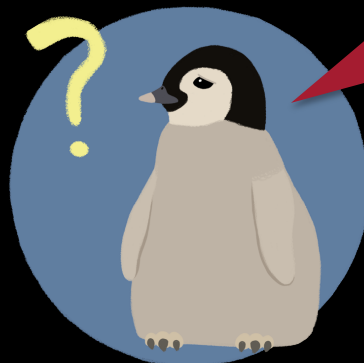
Prompt

GPT-2

Model

[illegible]

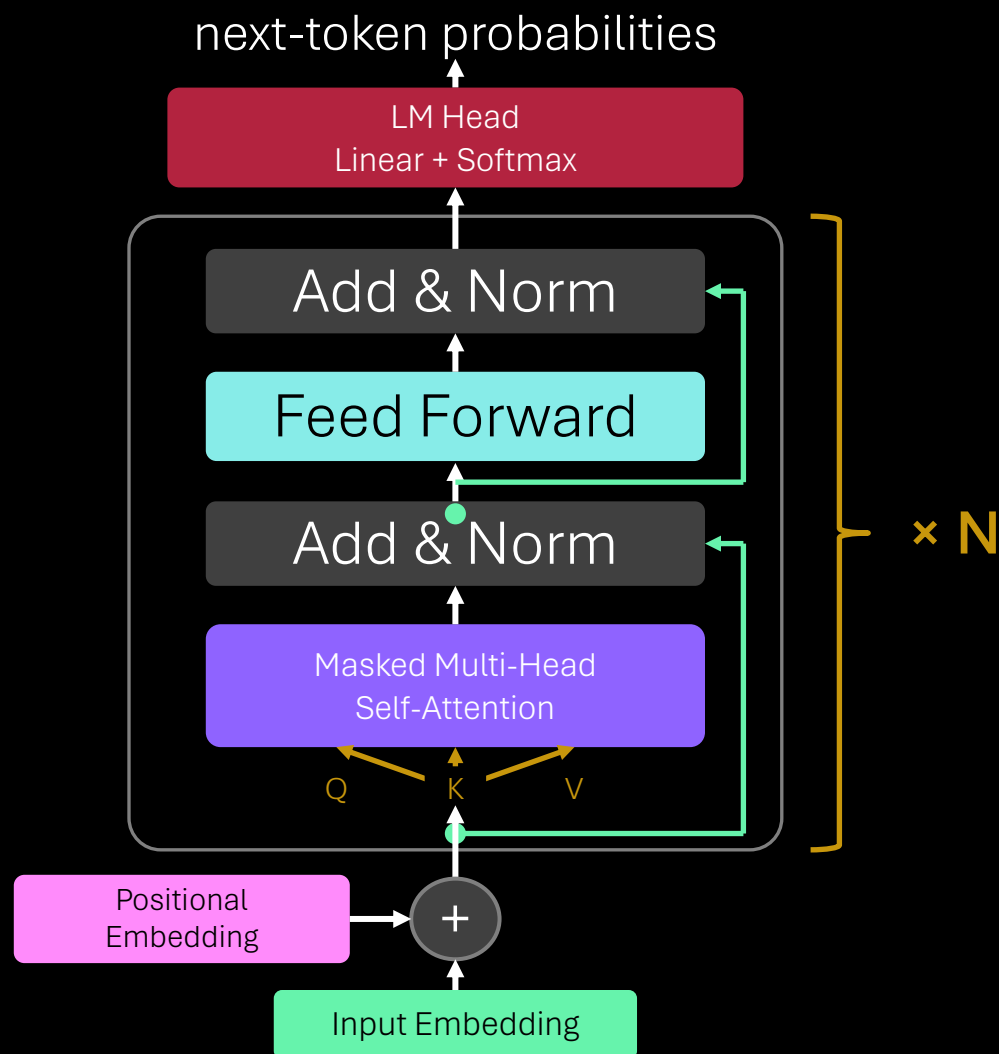
Response



Why is it giving such gibberish at the end?

Transformers

Generative Pre-Trained Transformer (GPT)



A Vanilla GPT like GPT-2 can struggle with:

- Following **specific instructions**
- Providing **factual** or **relevant** answers
- Maintaining **clarity** over **longer** interactions

Outline

- › Transformers (Recap)
- › **Instruction-Tuned LLMs (How Does ChatGPT Work?)**
 - › Instruction Training Overview
 - › RLHF (Reinforcement Learning from Human Feedback)
 - › PPO (Proximal Policy Optimization)
 - › Clipping

Instruction-Tuned LLMs

What are Instruction-Tuned LLMs?

- Instruction-tuned LLMs are large language models fine-tuned to better understand and follow human instructions.
- This tuning improves their ability to generate useful, relevant, and coherent responses, making them more effective at tasks such as answering questions, providing explanations, and performing specific commands.

Outline

- › Transformers (Recap)
- › **Instruction-Tuned LLMs (How Does ChatGPT Work?)**
 - › **Instruction Training Overview**
 - › RLHF (Reinforcement Learning from Human Feedback)
 - › PPO (Proximal Policy Optimization)
 - › Clipping

Instruction Training | GPT-3.5 Turbo and Beyond

- › GPT models are trained to **predict the next word** in a sentence given the context of the preceding words.
- › The model does not have access to the specific instructions or intentions of the user. Therefore, it may not always **align answers with what the user wants**.
- › To bridge this gap, newer models go through additional training steps.



Instruction Training

SFT

Supervised Fine-Tuning trains the model on example **instructions and responses** so it learns to **follow user instructions** more accurately.

RLHF

Reinforcement Learning from Human Feedback (RLHF) is used to incorporate human feedback into the training process to better align the model's output with **user intent**.

GPT Instruction Tuning

Stage	Pre-Training (GPT)	 Supervised fine-tuning	 Reward Modeling	 Reinforcement Learning
Data	Raw Internet	Assistant Responses	Response Comparisons	Sample Prompts
Algorithm	Language Model	Language Model	Binary Classification	Reinforcement Learning
Output	Base Model 	SFT Model 	 Reward Model	RLHF Model 

Pretraining Stage | Example Model

Model Card

Vocabulary size: 50,257

Context length: 2048

Parameters: 175 Billion

Tokens trained: 300 Billion

Time: ~1 month of training

Resources: ~1000-10000 GPUs

Cost: ~10 Million USD

Year: 2020

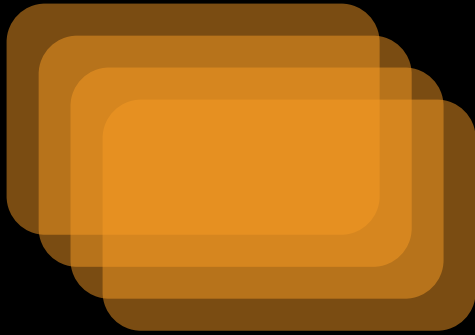
GPT-3



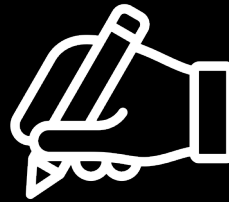
Instruction Training

Supervised Fine-Tuning Model

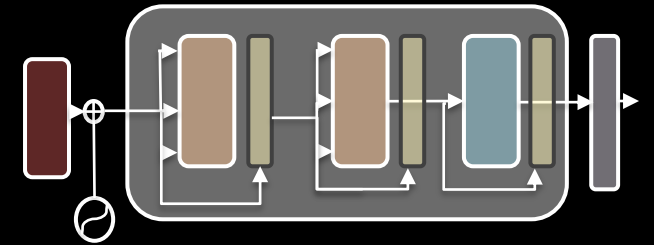
The dataset is a large online collection of **math solutions** and **reasoning examples**, including both **correct** and **incorrect answers**, as well as a variety of **viewpoints** and **beliefs**.



Prompt sampled from dataset.



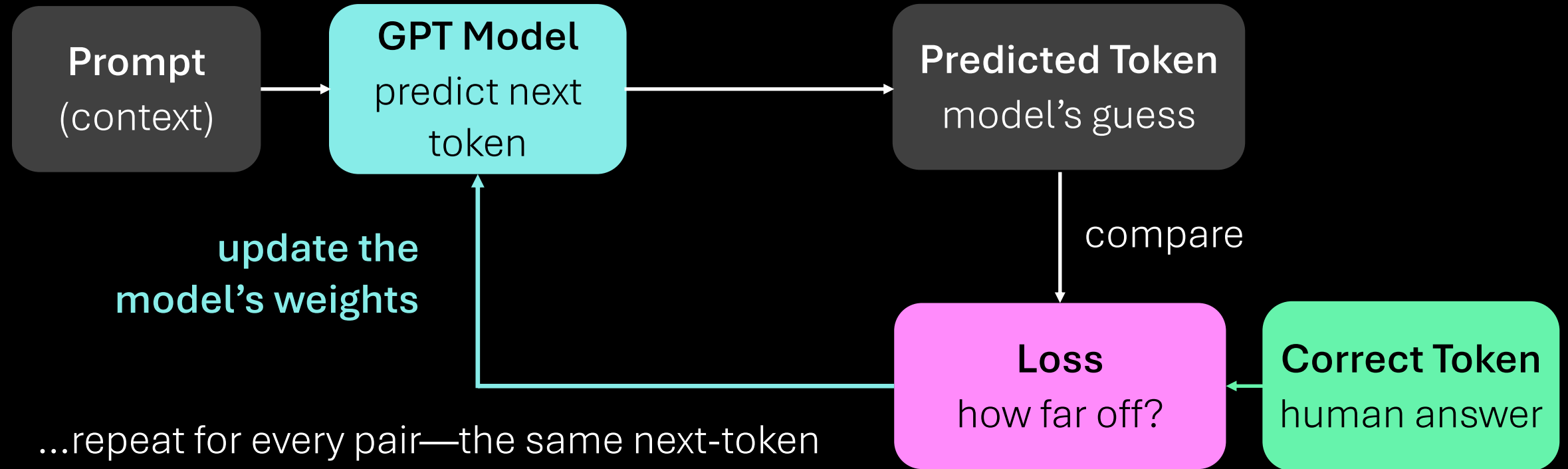
40 contractors wrote the desired output behavior.



Input/Output pairs are used to fine-tune GPT with supervised learning.

Instruction Training

What fine-tuning actually does:



...repeat for every pair—the same next-token training as pre-training, now on curated answers.

How to Train Your Dog AI

SFT leads to a model that:

- › Follows **basic instructions**
- › Generates **decent responses**

But there's a problem...

SFT Model



I need a concise summary of the key health benefits of regular exercise. Base your summary on the understanding that exercise improves cardiovascular health, strengthens bones, and boosts mental well-being.

Your summary must be a single paragraph. Please make sure to cover all three points.

Important Final Instruction: Do not use the letter 'e' anywhere in your summary.

SFT

“A daily workout grants a ton of good things for your body. A strong heart, tough structural supports, and a happy mind all follow a good physical program. Your body will work right and your brain will find joy. This is a vital part of a long, satisfactory human span.”

The SFT model's rigid adherence to the constraints completely destroyed the quality and utility of the summary.

How to Train Your Dog AI

SFT leads to a model that:

- › Follows **basic instructions**
- › Generates **decent responses**

But there's a problem...



- › SFT only learns from **fixed** prompt-response pairs
- › It can't go beyond what it's seen
- › Doesn't know which response is **better** when there are multiple valid options
- › We can't **label** everything **forever**

So, what do we do?

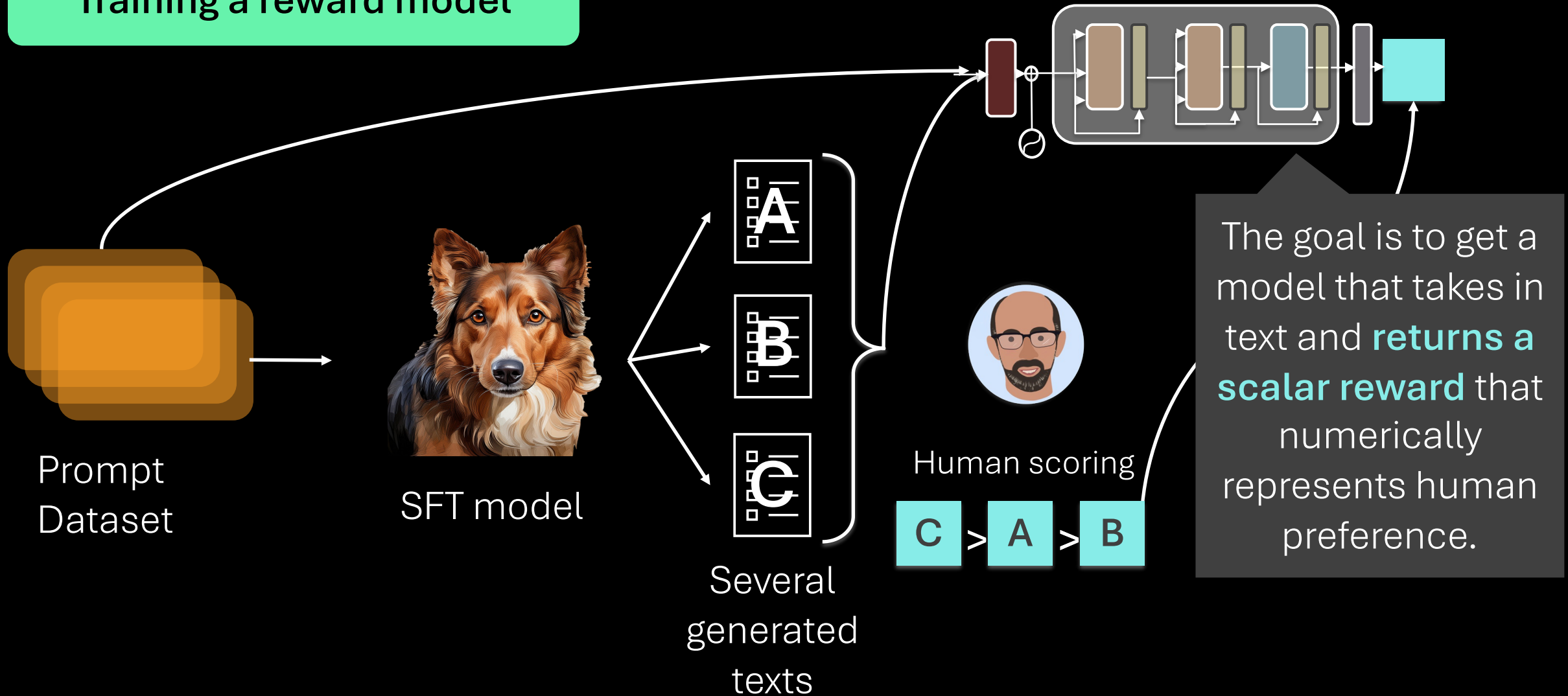


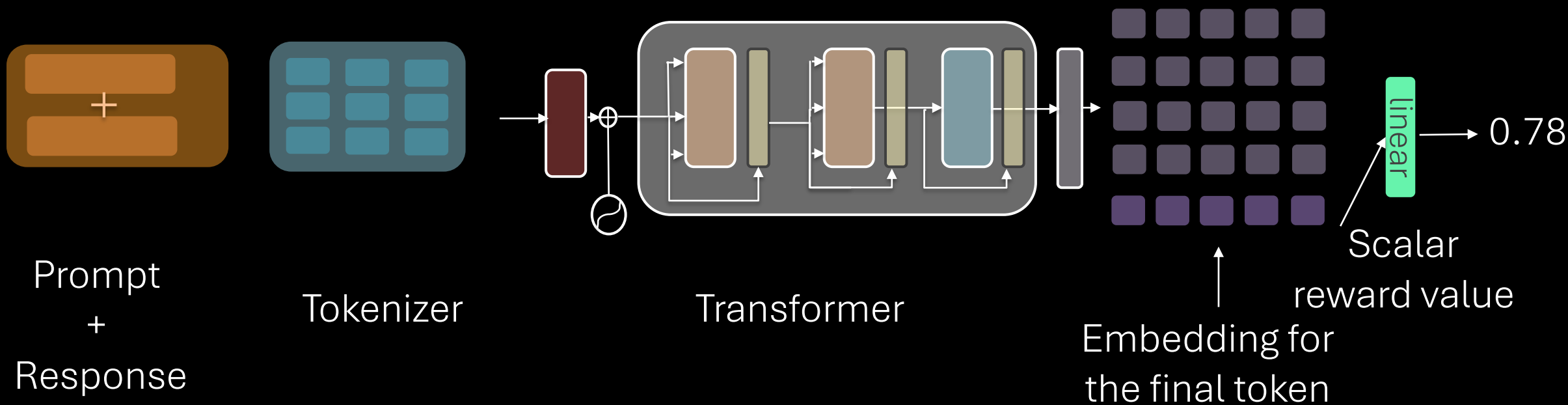
Outline

- › Transformers (Recap)
- › Instruction-Tuned LLMs (How Does ChatGPT Work?)
 - › Instruction Training Overview
 - › **RLHF (Reinforcement Learning from Human Feedback)**
 - › PPO (Proximal Policy Optimization)
 - › Clipping

RLHF | Training a Reward Model

Training a reward model

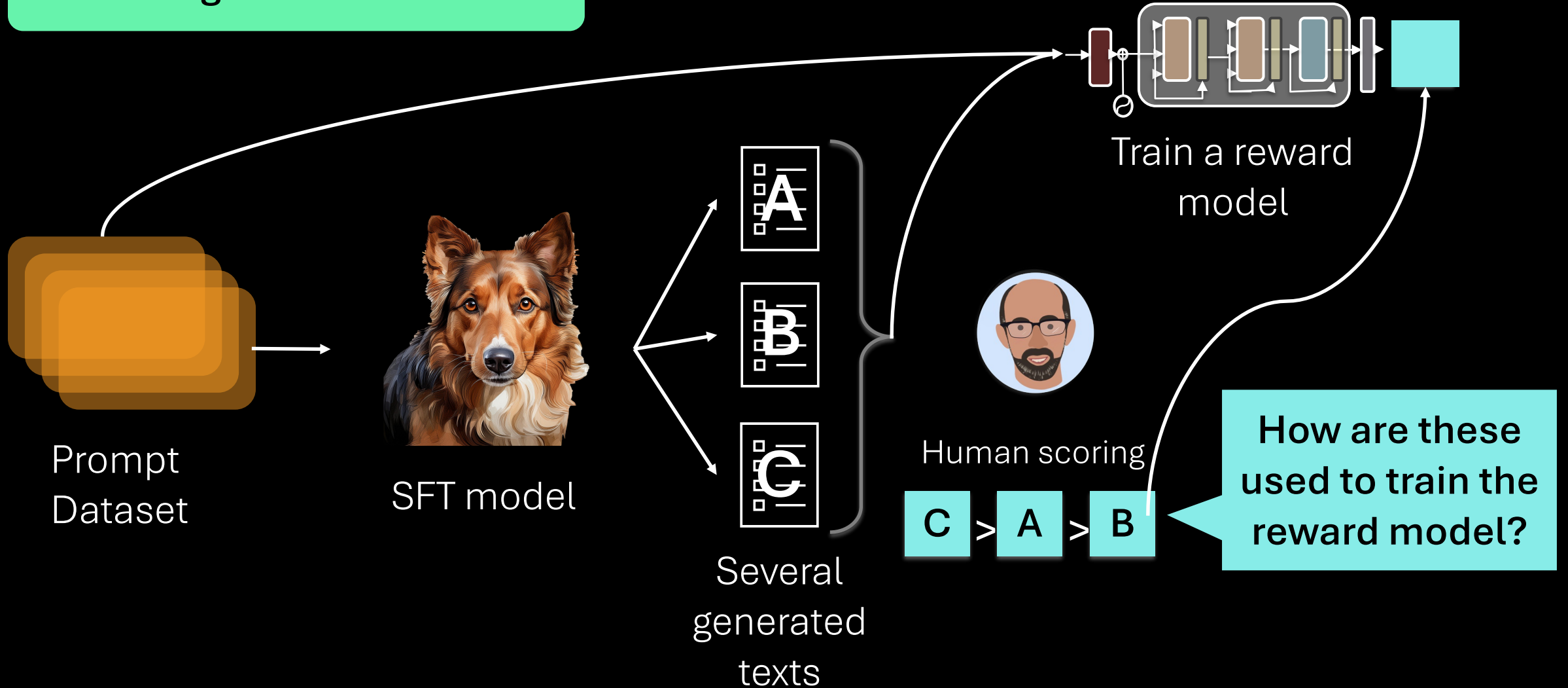




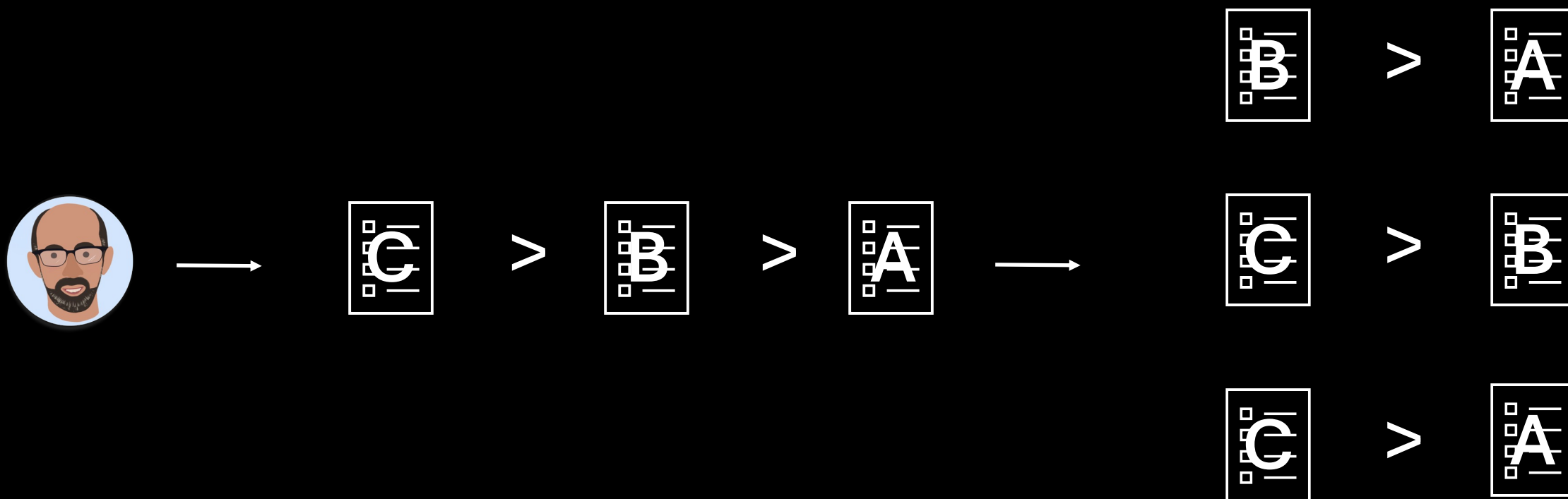
The model is trained on pairs of responses, adjusting its weights so that the scalar reward for a human-preferred response is higher than that for a less-preferred one.

RLHF | Training a Reward Model

Training a reward model



RLHF | Training a Reward Model



RLHF | Training a Reward Model

The intuition

A human compares two responses to the same prompt and picks the one they prefer.

Response i

preferred reward model gives it
the higher score ↑

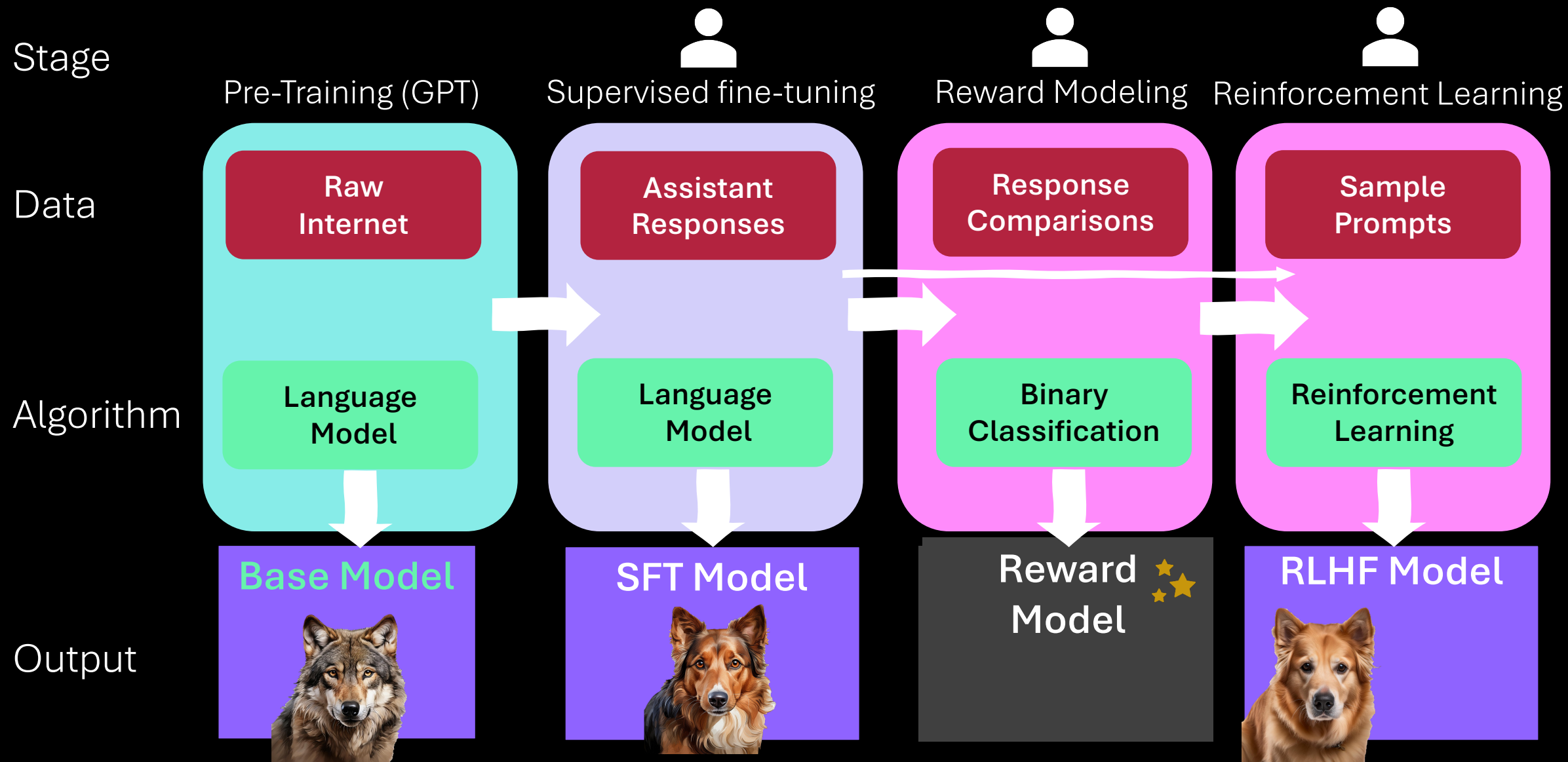
Response j

gets the lower score ↓

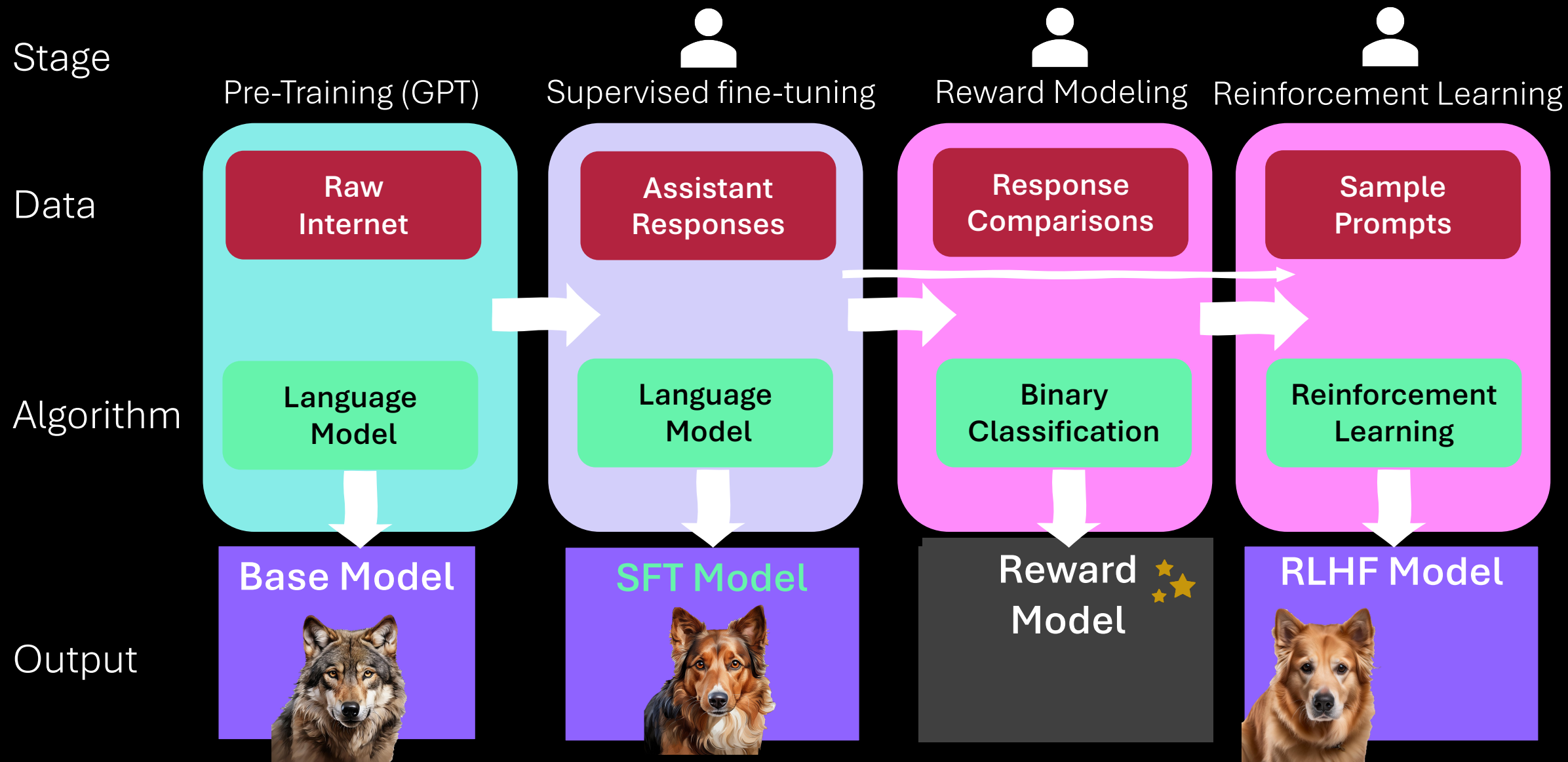
The whole idea: reward the preferred answer—push its score up and the other one down.

—.

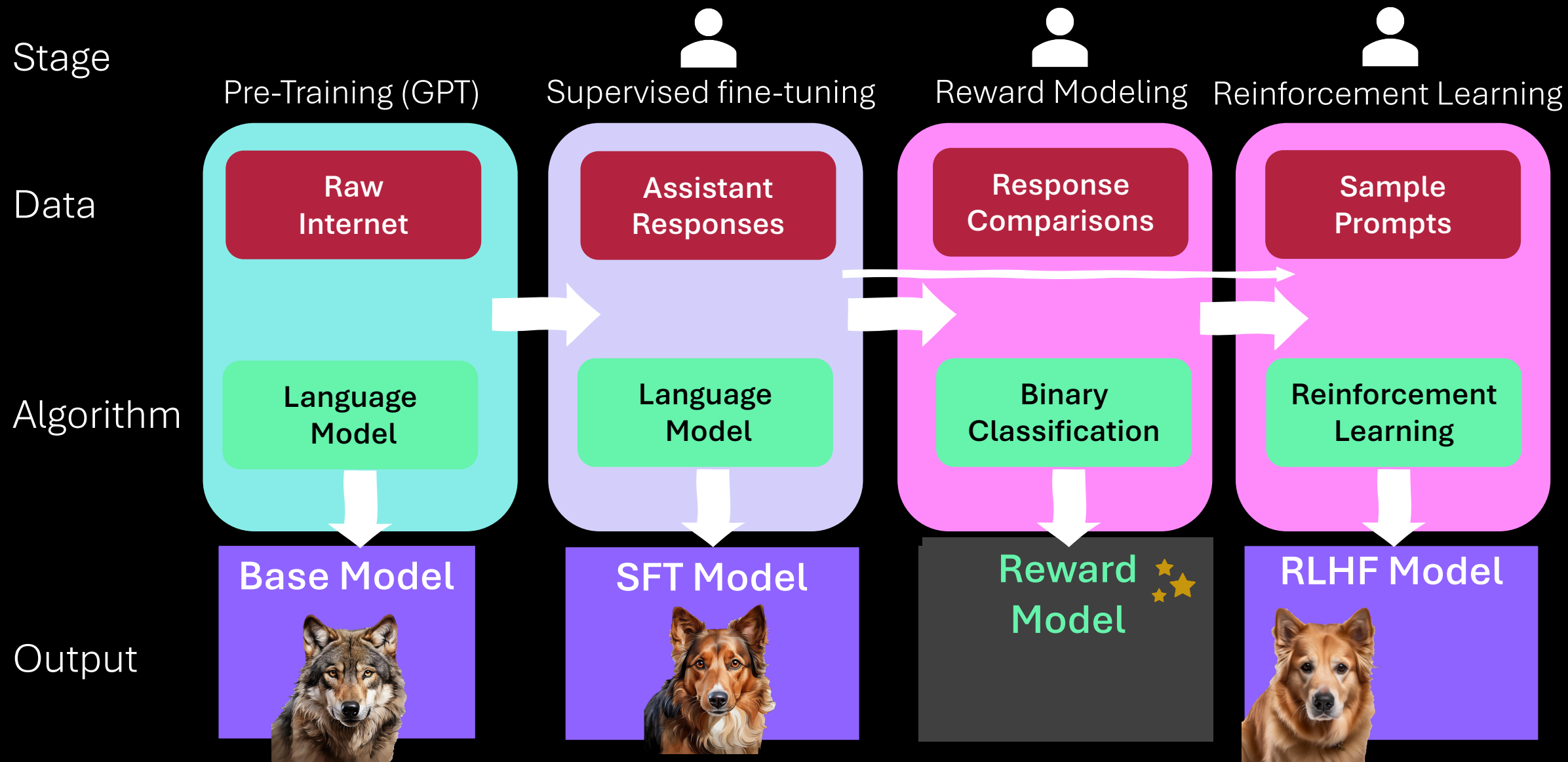
GPT Instruction Tuning



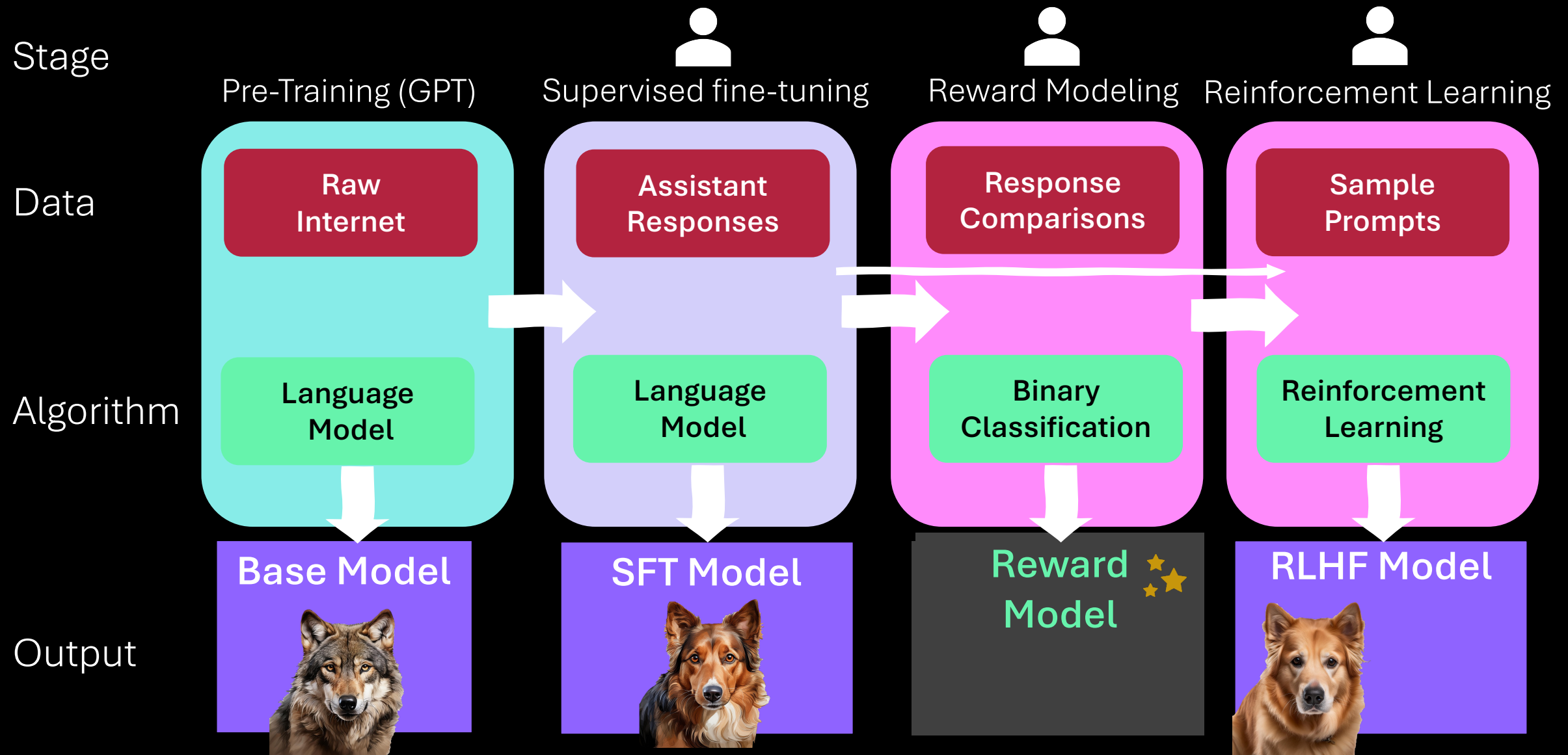
GPT Instruction Tuning



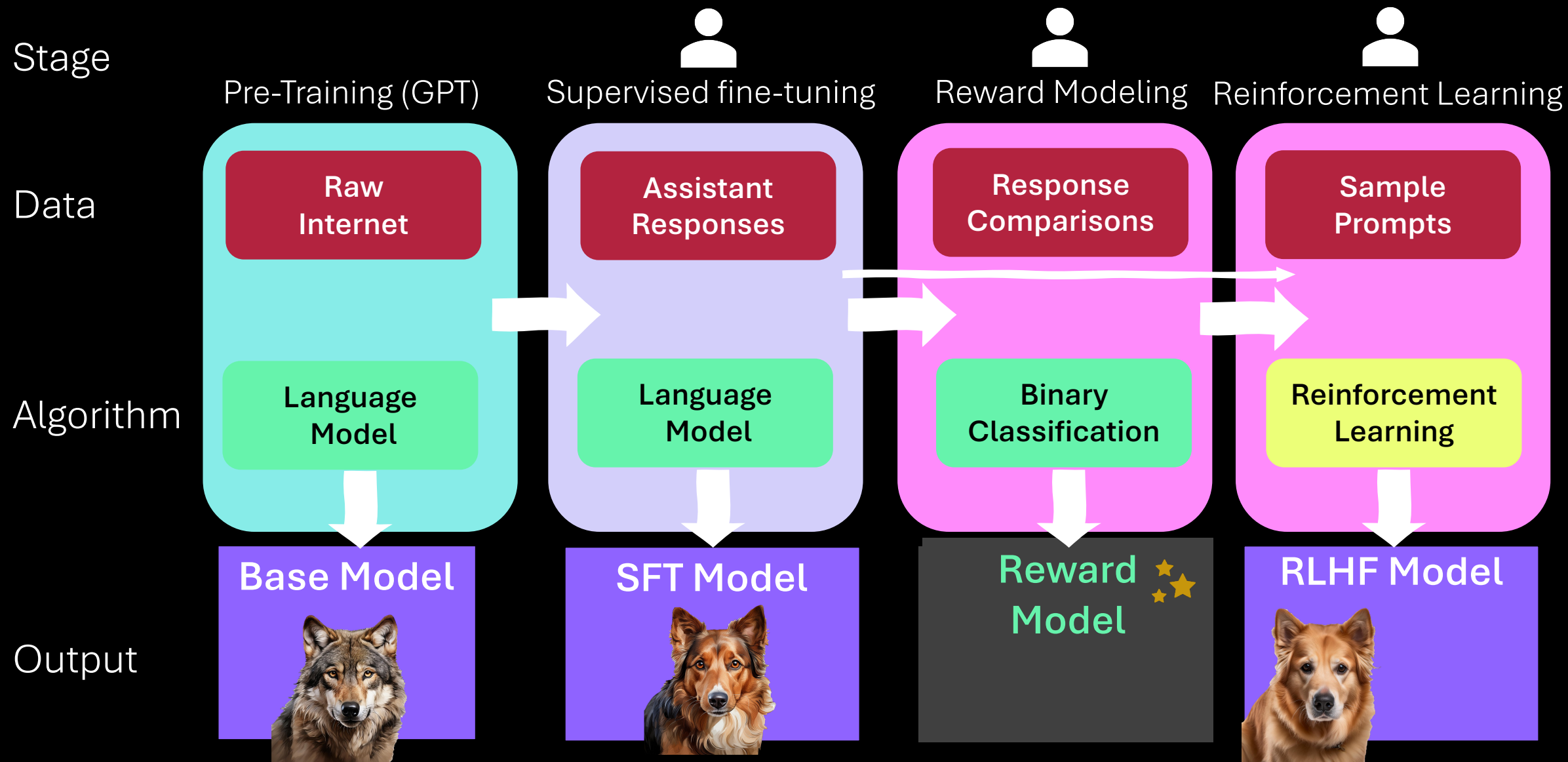
GPT Instruction Tuning



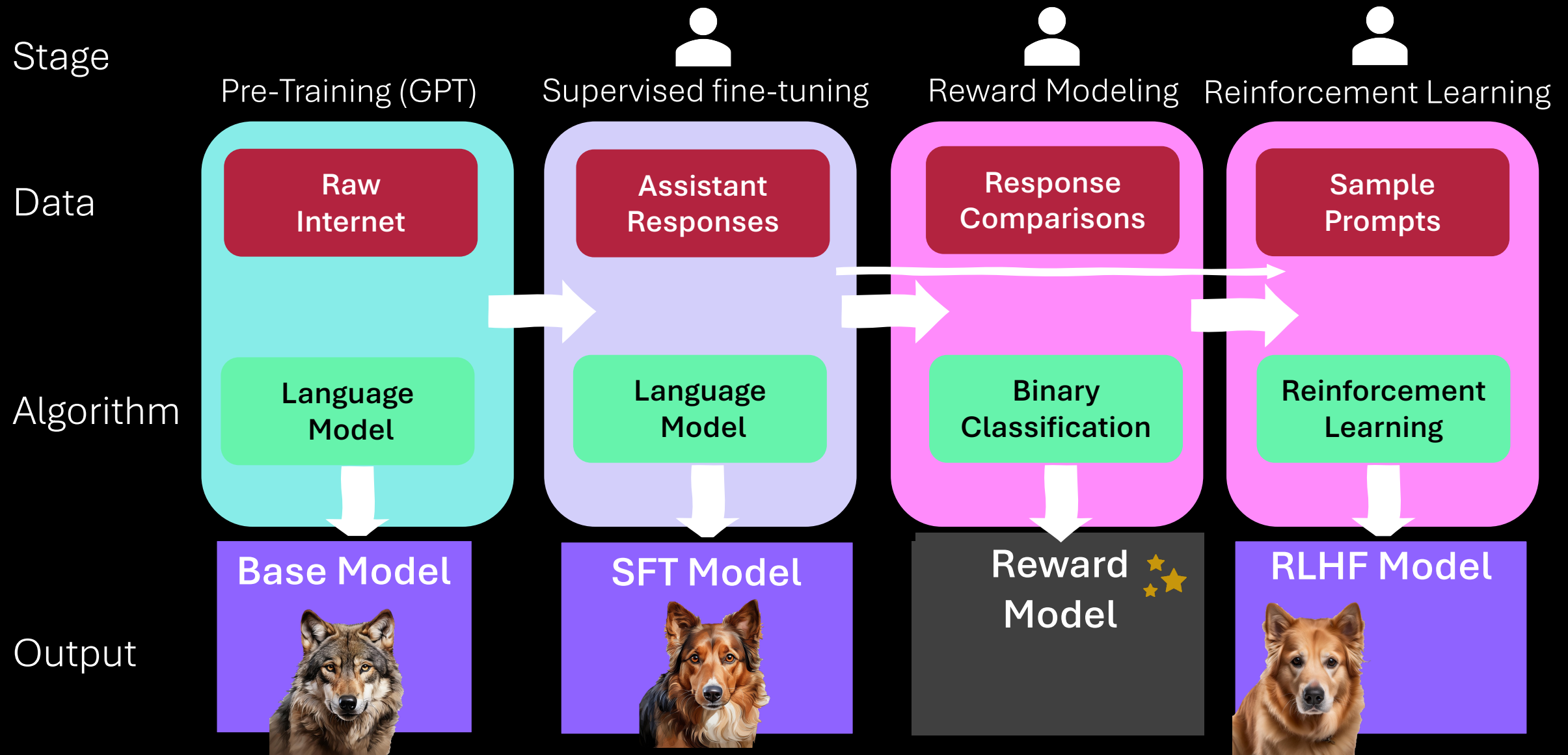
GPT Instruction Tuning



GPT Instruction Tuning



GPT Instruction Tuning



Instruction Training

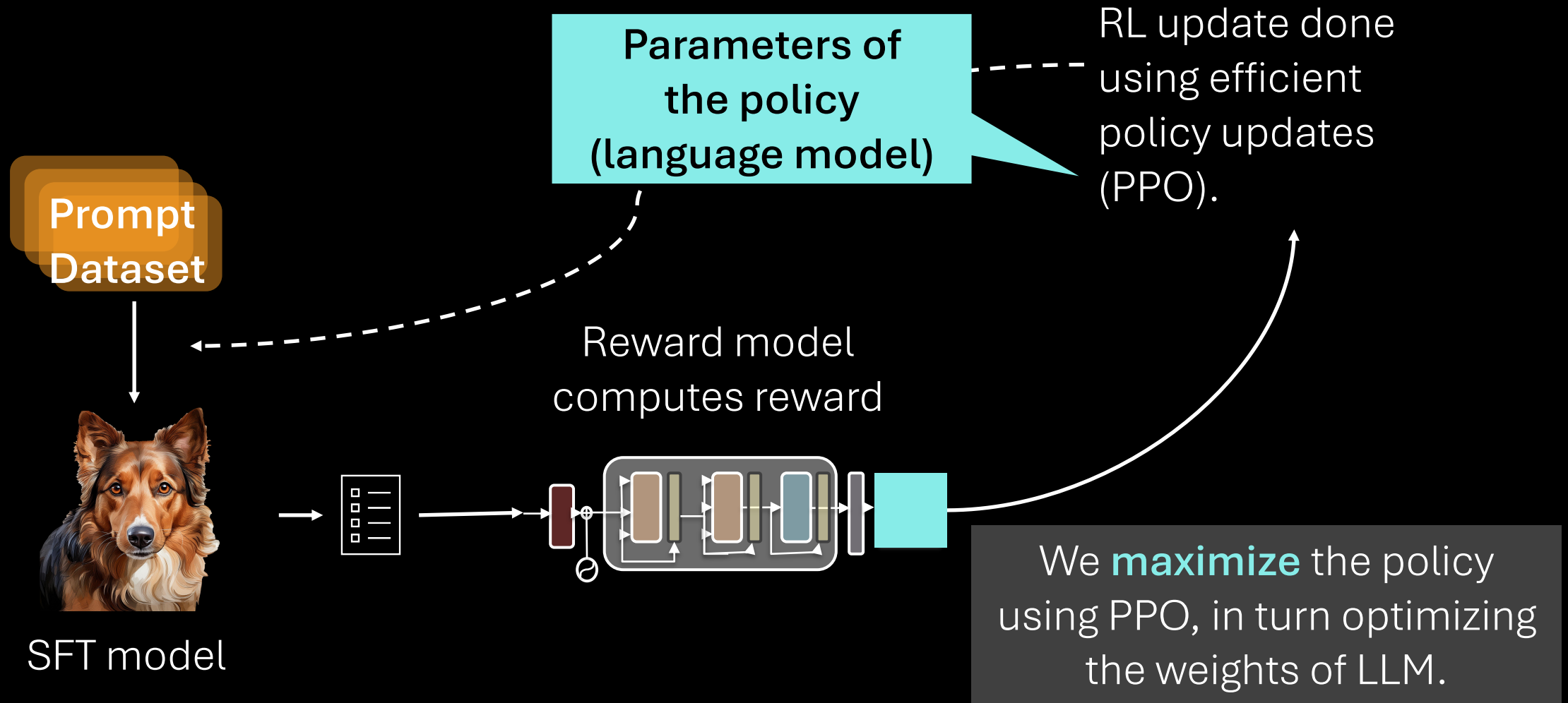
Let's frame this fine-tuning step as a reinforcement learning problem:

- › **Action space:** Each token in the vocabulary is an action. The model generates responses one token at a time.
- › **Policy:** The language model acts as the policy—it defines a probability distribution over the next token, given the prompt and previous tokens.
- › **Reward (critic):** The reward model is the critic—it scores full responses based on human preferences and provides feedback to improve the actor.

Outline

- › Transformers (Recap)
- › Instruction-Tuned LLMs (How Does ChatGPT Work?)
 - › Instruction Training Overview
 - › RLHF (Reinforcement Learning from Human Feedback)
 - › **PPO (Proximal Policy Optimization)**
 - › Clipping

RLHF | The Reinforcement Step | PPO



Source: PPO Algorithm

The Reinforcement Step | PPO

What is PPO?



- **Proximal Policy Optimization (PPO)** is a **policy gradient method** designed to improve a policy while preventing **large, destabilizing changes** between updates.
- Traditional policy gradient algorithms can make **big jumps** in **parameter space**, often **harming performance**.
- PPO achieves stability with a **simple approach**.

RLHF Model Training | No Clipping

Prompt example: “A poem about dogs”

“reward hacking” 

SFT Model

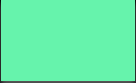


Prompt	...	Response 1	Dog poem dog poem dog poem
Prompt	...	Response 2	A dog is a pet.
Prompt	...	Response 3	A loyal friend, a happy face, a wagging tail all over the place.

< reward > = 5
< reward > = 0.5
< reward > = 3

RLHF Model



-  Prompt tokens
-  Completion tokens (by SFT model)
-  Reward token (by Reward Model)


Reward Model

RLHF Model Training | With Clipping

Prompt example: “A poem about dogs”

SFT Model



Prompt	...	Response 1	Dog poem dog poem dog poem	< reward > = 1.5
Prompt	...	Response 2	A dog is a pet.	< reward > = 0.5
Prompt	...	Response 3	A loyal friend, a happy face, a wagging tail all over the place.	< reward > = 3

RLHF Model



-  Prompt tokens
-  Completion tokens (by SFT model)
-  Reward token (by Reward Model)

Reward Model



GPT Instruction Tuning

