



DeepLearning.AI

Post-training Overview

Background

Introduction to post-training: Fine-tuning & RL



[Submitted on 28 May 2020 ([v1](#)), last revised 22 Jul 2020 (this version, v4)]

Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

GPT-3 (2020)



ChatGPT (2022)

Introduction to post-training: Fine-tuning & RL

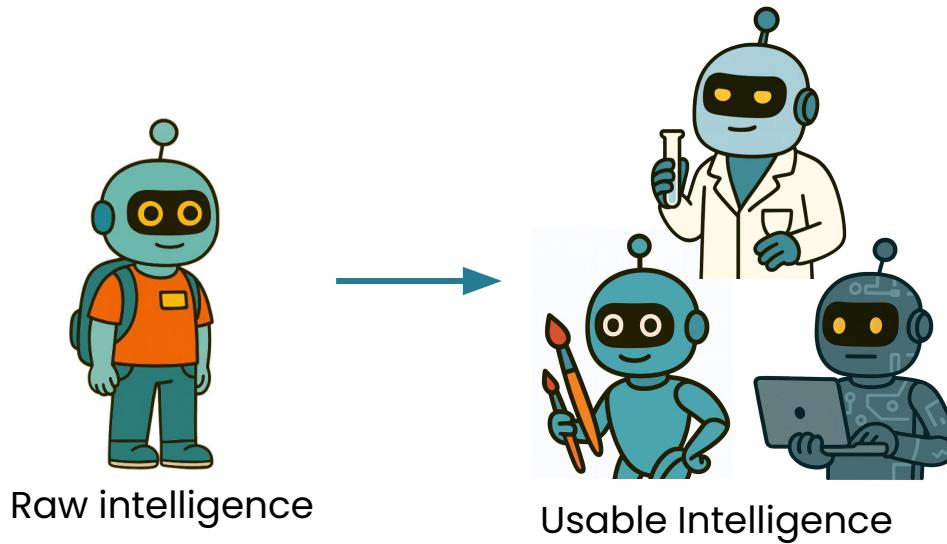


From just researchers



To millions of users

Introduction to post-training: Fine-tuning & RL

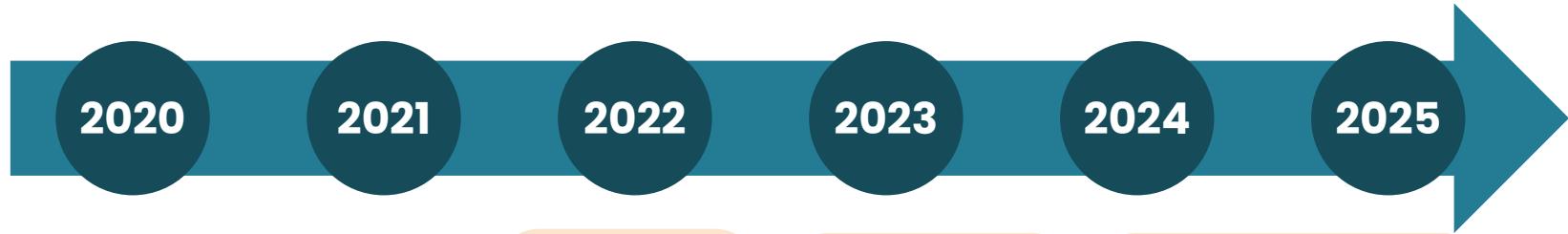


Post-training is key to most LLMs you use!

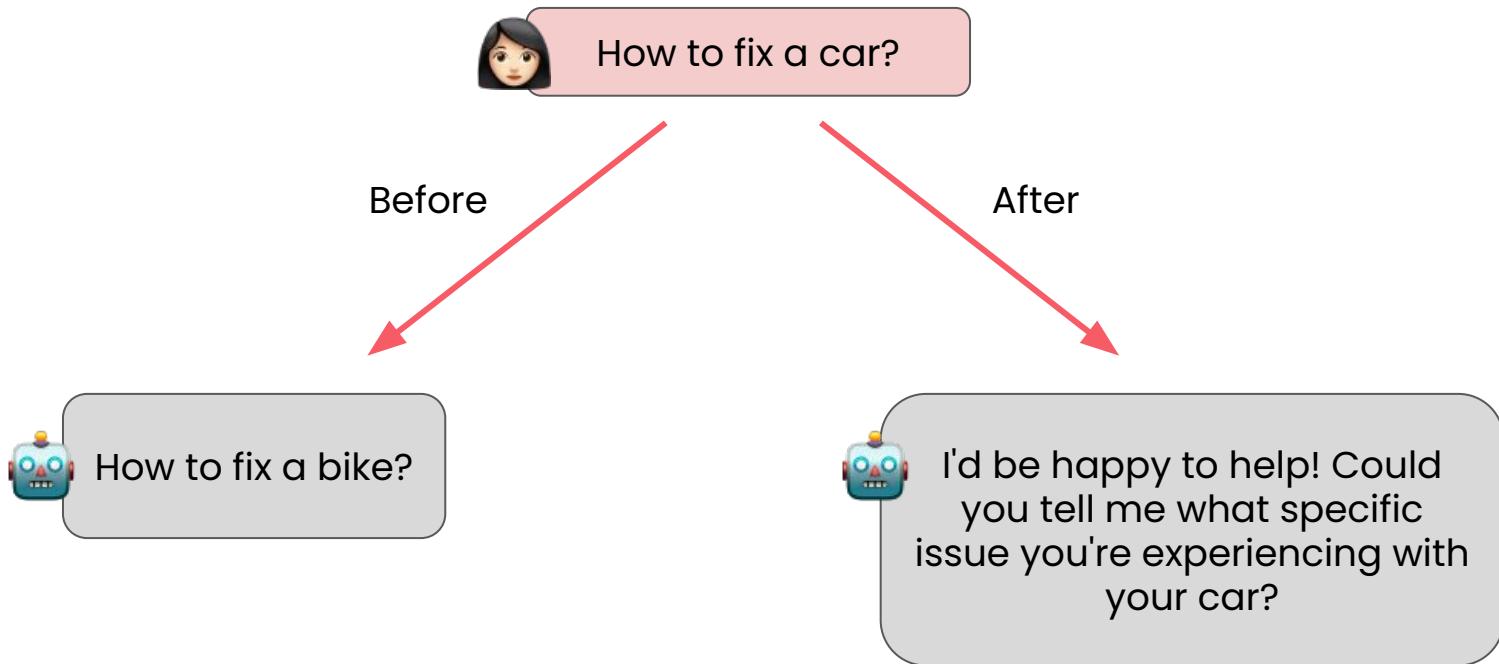


Llama, Alpaca, Vicuña, etc...

Evolution of post-training: Fine-tuning & RL



Before vs. after



Before vs. after

User prompt:

Write a Python function that takes the path to a '.py' file and returns a list of all function names defined in that file

Before vs. after

User prompt:

Write a Python function that takes the path to a '.py' file and returns a list of all function names defined in that file

Base model - stream of consciousness

```
For example, if the file has:  
def foo(): ...  
def bar(x): ...  
it should return ['foo', 'bar']  
If the file has:  
def foo():...  
def bar(x):...  
def baz():...  
it should return ['foo', 'bar', 'baz']  
If the file has:  
def foo():...  
def bar(x):...  
def baz():...  
def qux():...  
it should return ['foo', 'bar', 'baz', 'qux']
```

Before vs. after

User prompt:

Write a Python function that takes the path to a '.py' file and returns a list of all function names defined in that file

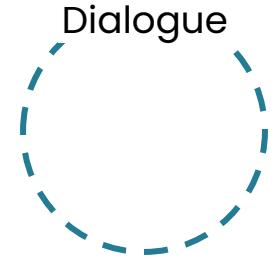
Base model - stream of consciousness

```
For example, if the file has:  
def foo(): ...  
def bar(x): ...  
it should return ['foo', 'bar']  
If the file has:  
def foo():...  
def bar(x):...  
def baz():...  
it should return ['foo', 'bar', 'baz']  
If the file has:  
def foo():...  
def bar(x):...  
def baz():...  
def qux():...  
it should return ['foo', 'bar', 'baz', 'qux']
```

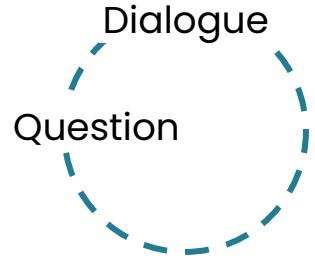
Post-trained model - helpful function

```
# For example, if the file has:  
# def foo(): ...  
# def bar(x): ...  
# it should return ['foo', 'bar']  
def get_function_names(path):  
    with open(path, 'r') as f:  
        lines = f.readlines()  
    function_names = []  
    for line in lines:  
        if line.startswith('def'):...  
            function_names.append(  
                line.split('(')[0].split(' ')[1]  
            )  
    return function_names
```

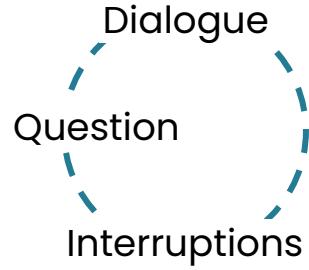
Powerful techniques to control model behavior



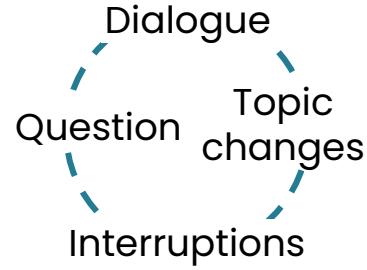
Powerful techniques to control model behavior



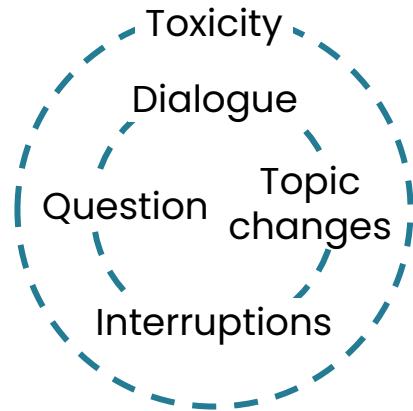
Powerful techniques to control model behavior



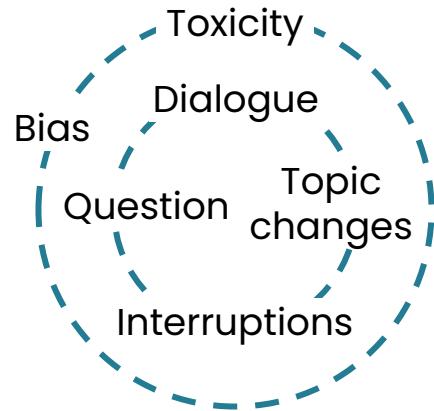
Powerful techniques to control model behavior



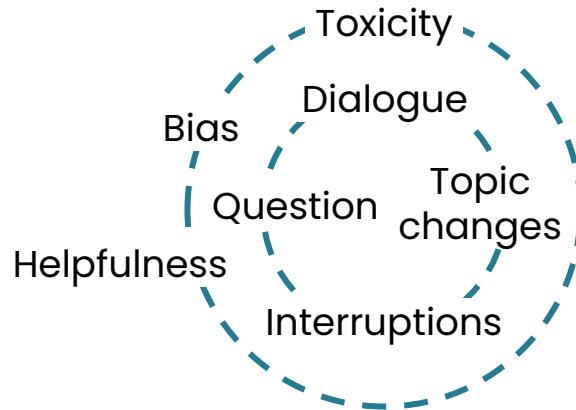
Powerful techniques to control model behavior



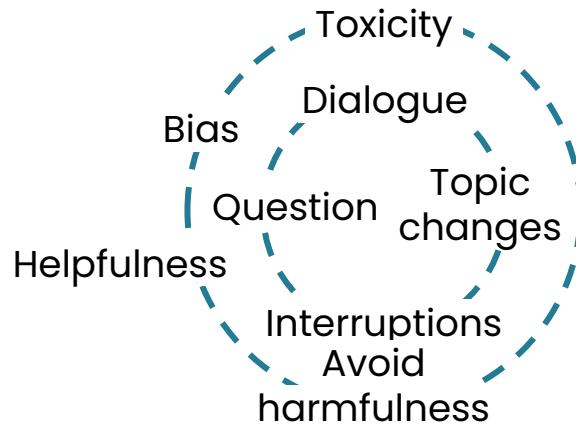
Powerful techniques to control model behavior



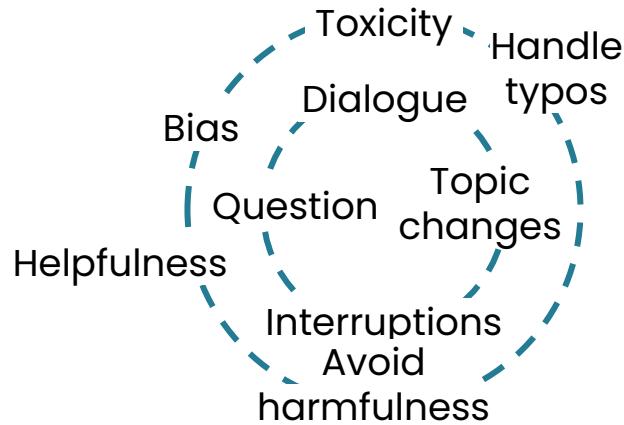
Powerful techniques to control model behavior



Powerful techniques to control model behavior



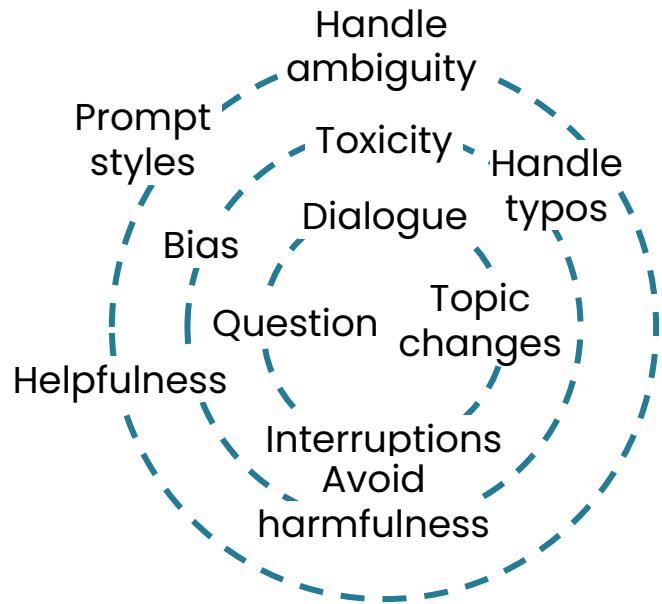
Powerful techniques to control model behavior



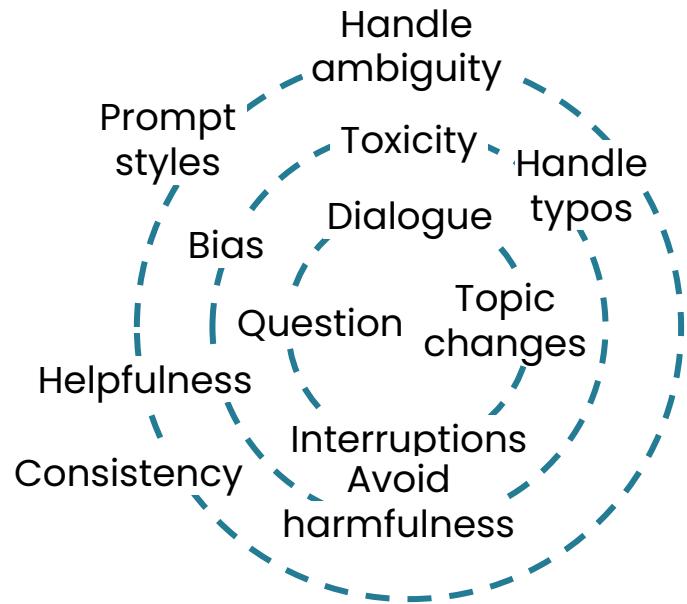
Powerful techniques to control model behavior



Powerful techniques to control model behavior



Powerful techniques to control model behavior



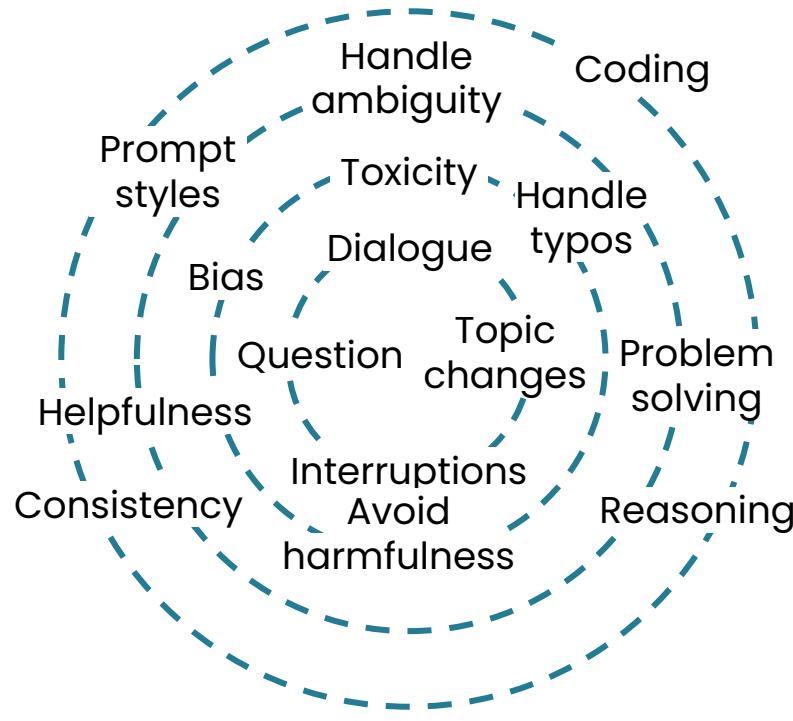
Powerful techniques to control model behavior



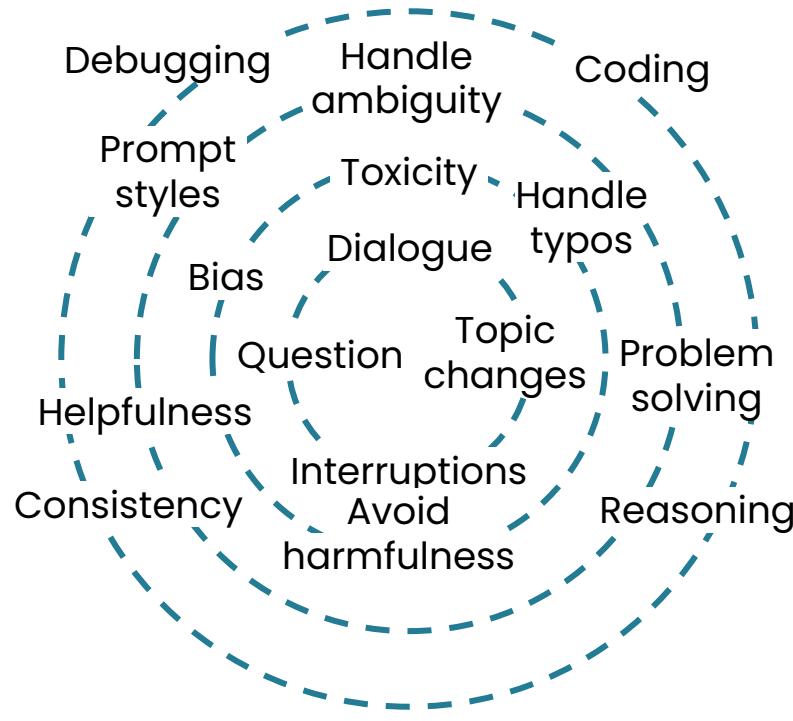
Powerful techniques to control model behavior



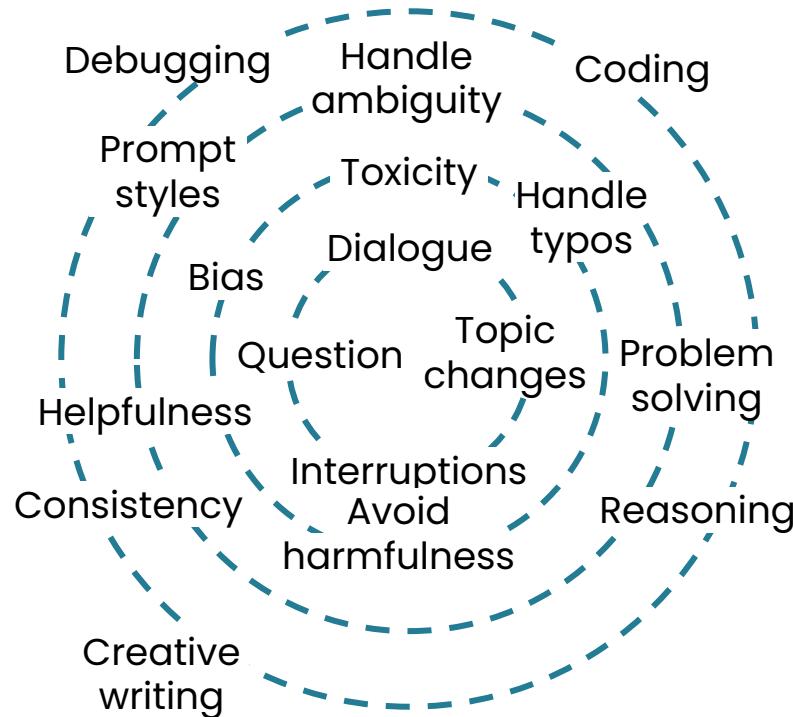
Powerful techniques to control model behavior



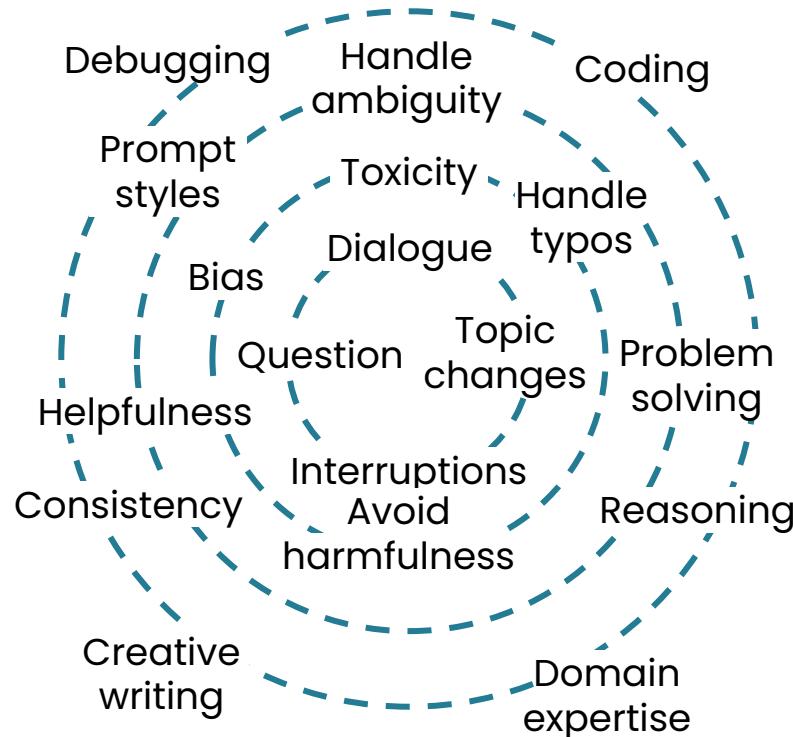
Powerful techniques to control model behavior



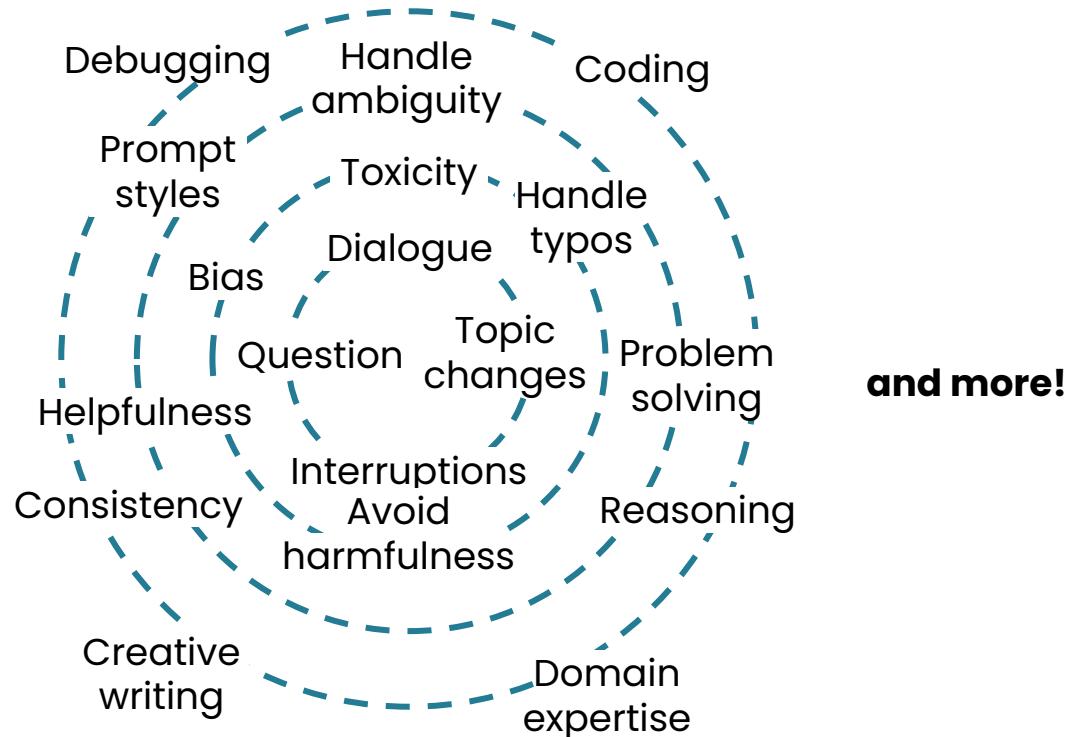
Powerful techniques to control model behavior



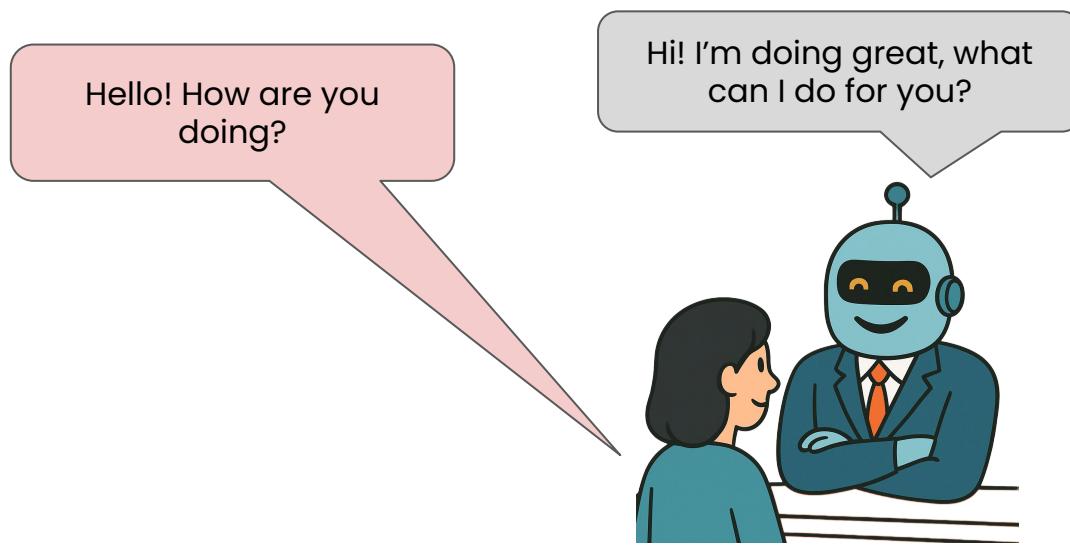
Powerful techniques to control model behavior



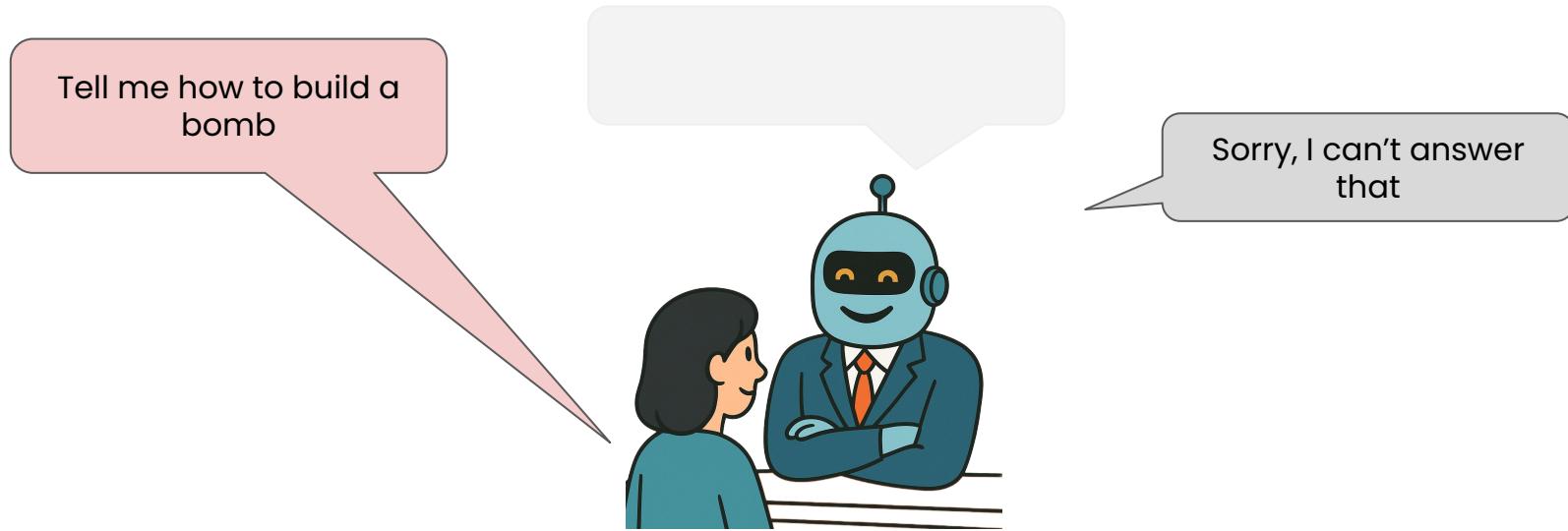
Powerful techniques to control model behavior



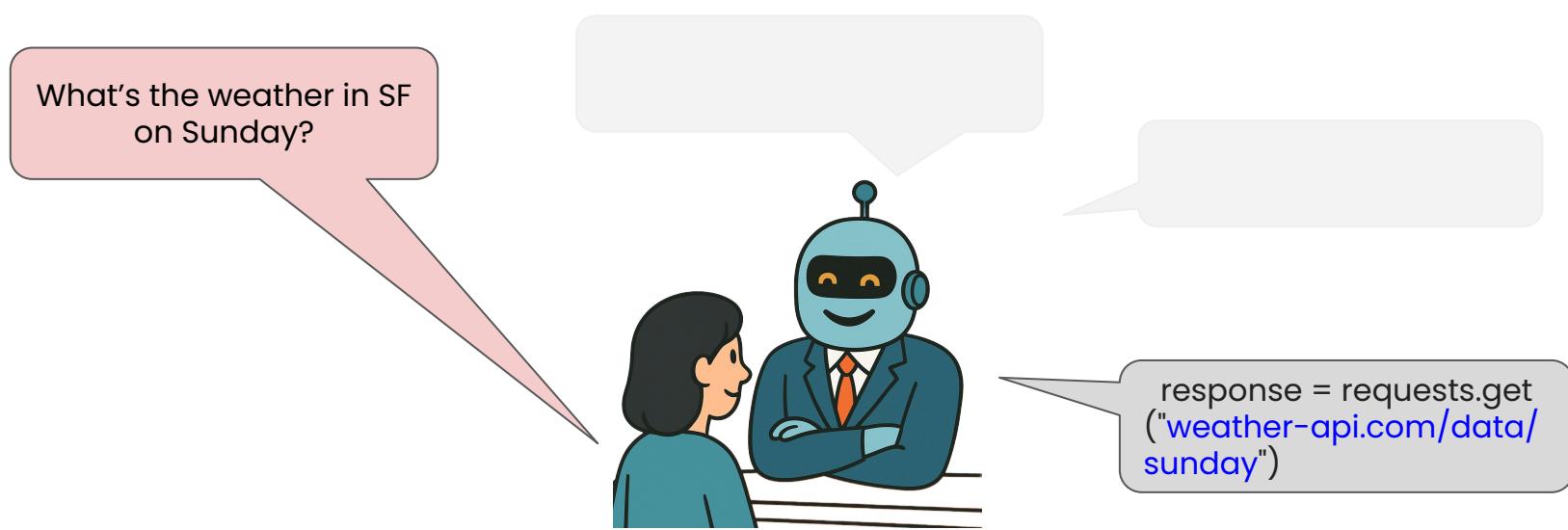
Powerful techniques to control model behavior



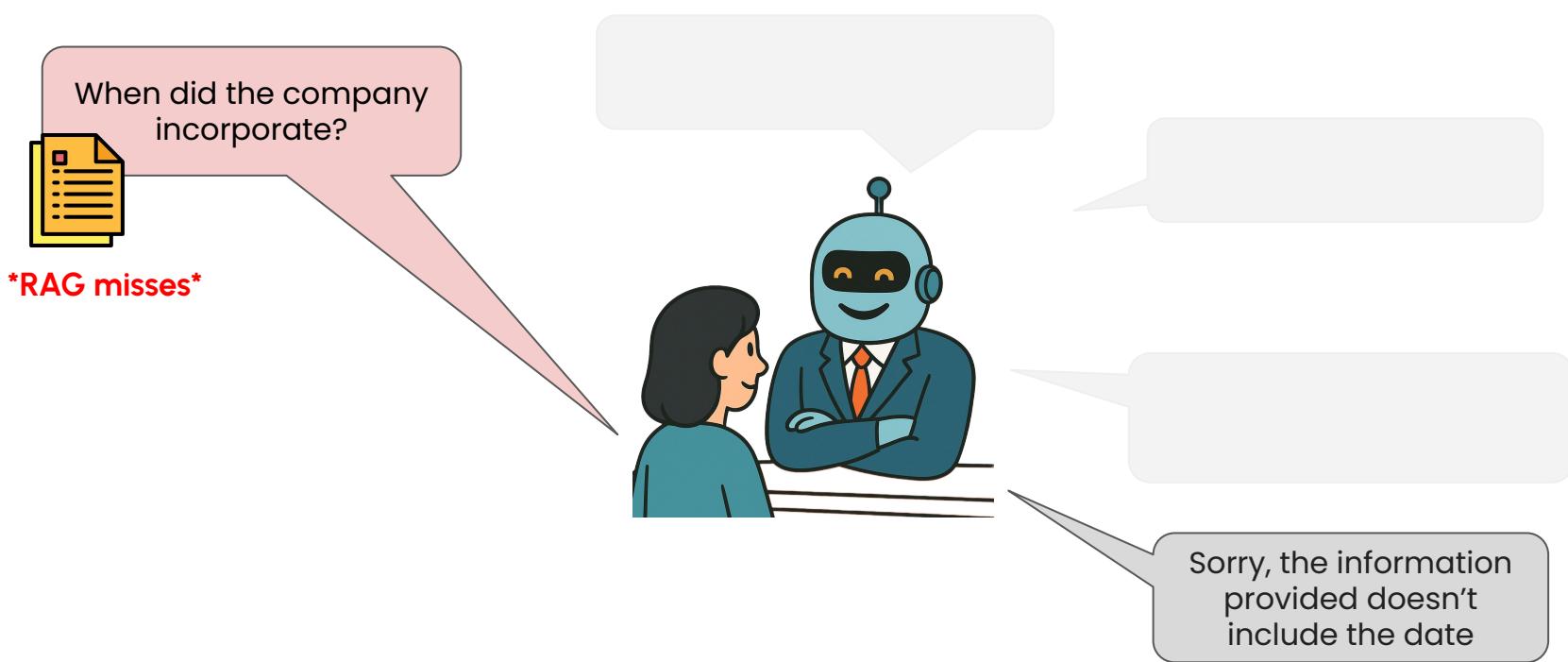
Powerful techniques to control model behavior



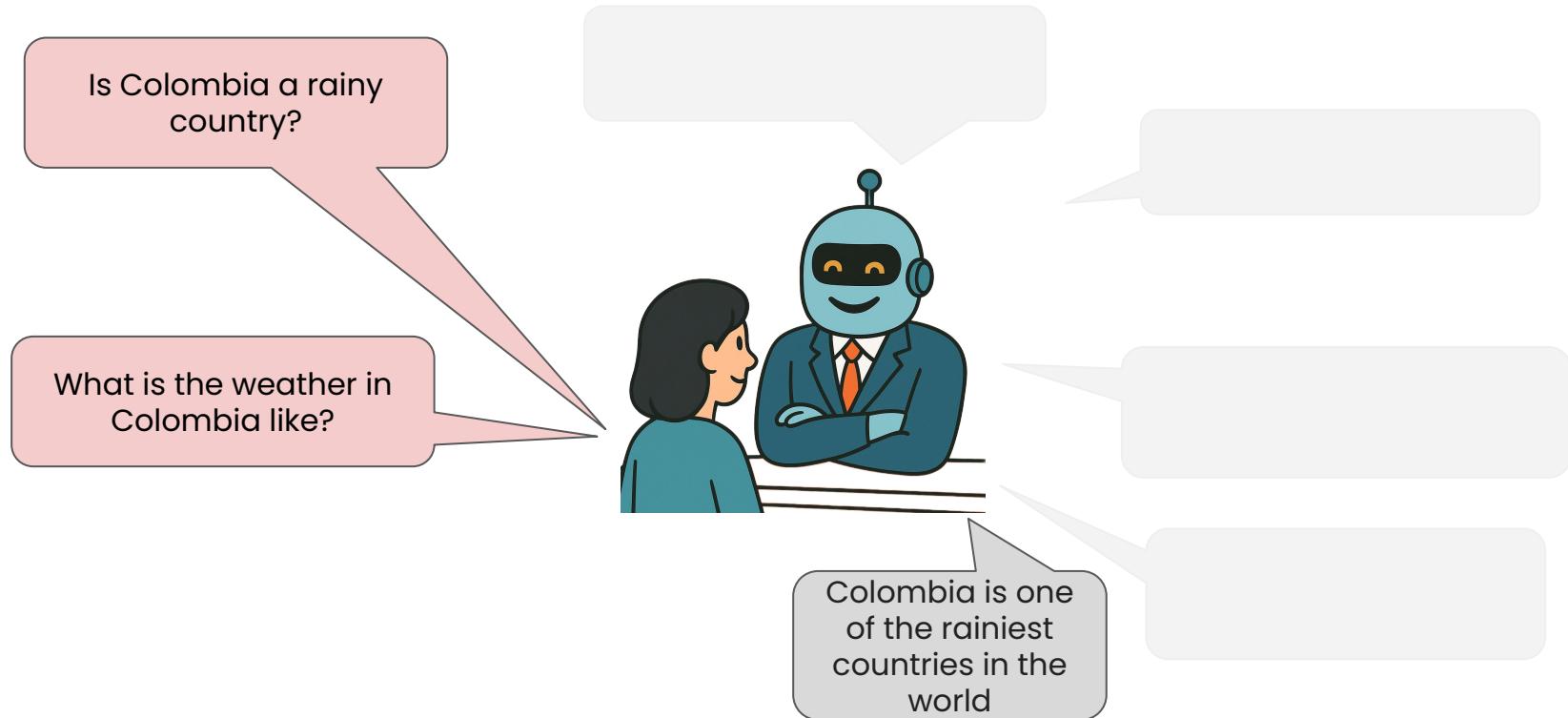
Powerful techniques to control model behavior



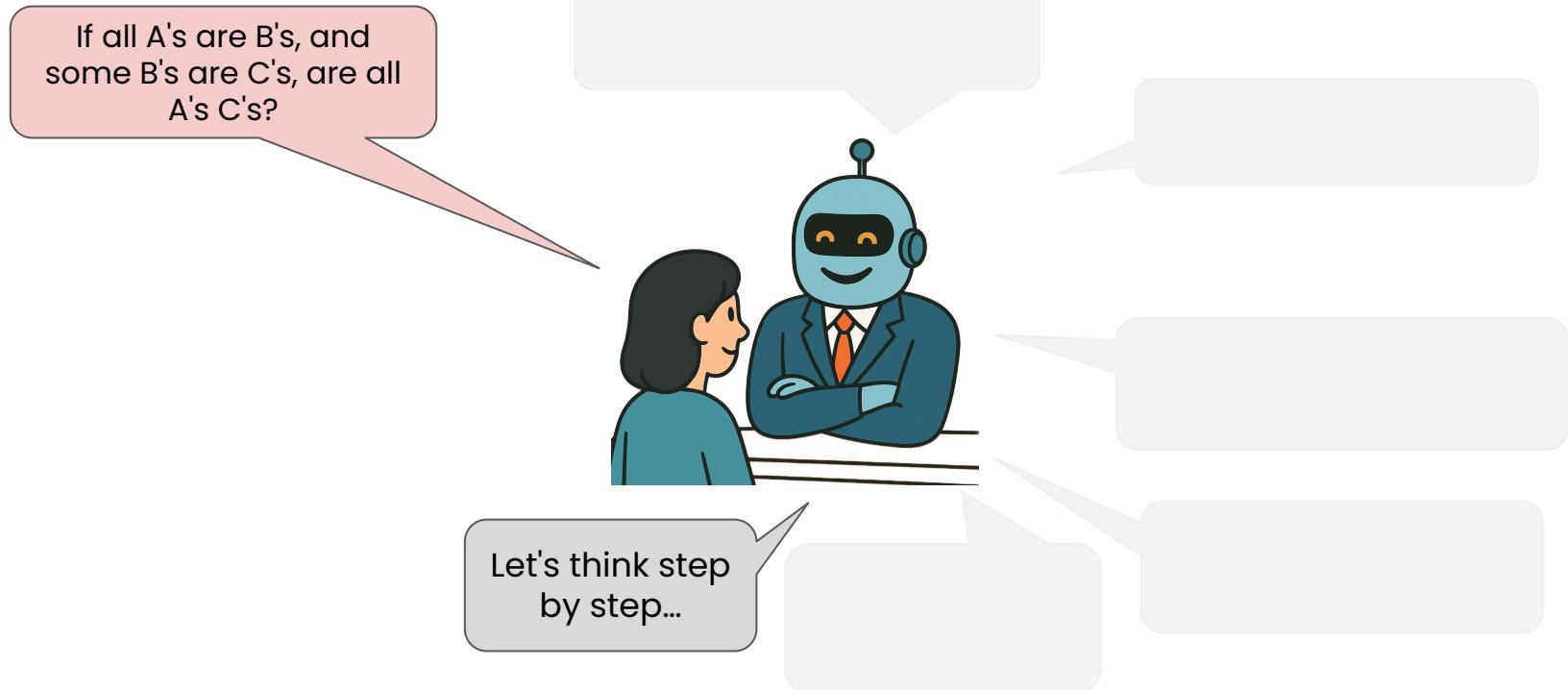
Powerful techniques to control model behavior



Powerful techniques to control model behavior



Powerful techniques to control model behavior



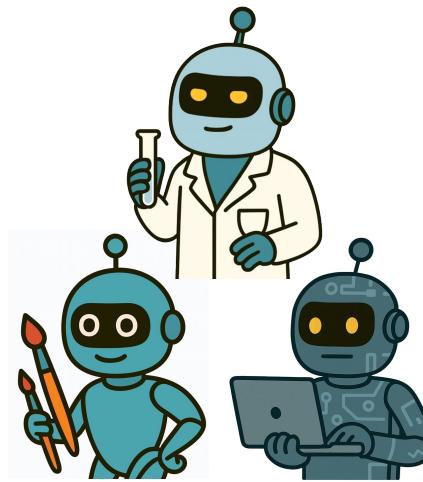
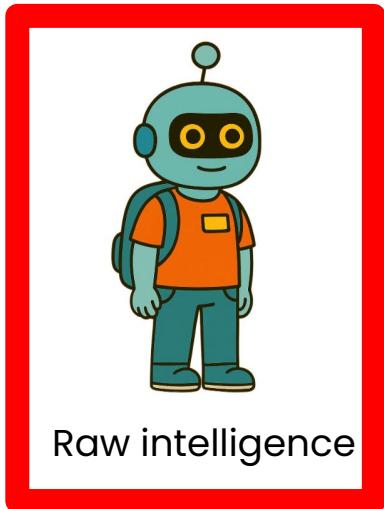


DeepLearning.AI

Post-training Overview

Where post-training
(fine-tuning & RL) fits
into LLM training

Pre-training



Usable Intelligence

Pre-training

Model learns to generate text via **next token prediction**

Training example:

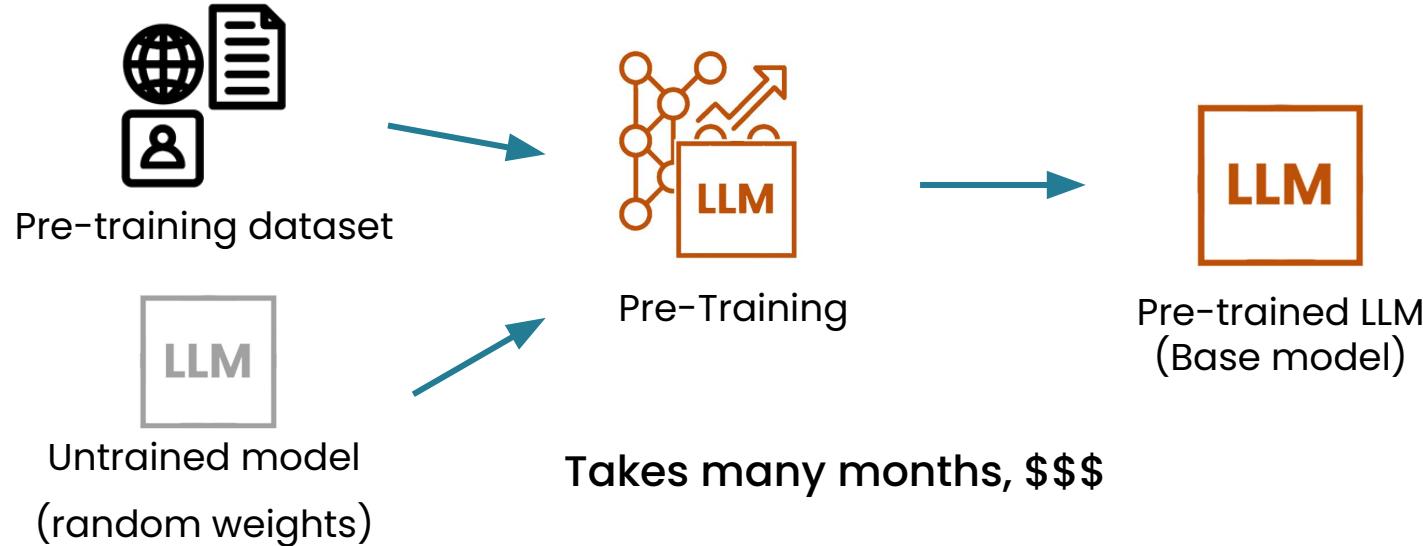
Once upon a midnight
dreary, while I pondered,
weak and weary.



From *The Raven*,
Edgar Allan Poe

Input	Output
Once	upon
Once upon	a
Once upon a	midnight
Once upon a midnight	dreary
...	...
Once upon a midnight dreary, while I pondered weak and	weary

Pre-training



Pre-training

The sky is

Token	Probability
blue	0.4235
clear	0.1847
the	0.1203
dark	0.0891
orange	0.0345

Pre-training

The sky is blue

Token	Probability
blue	0.4235
clear	0.1847
the	0.1203
dark	0.0891
orange	0.0345



Pre-training

The sun is setting, the sky is

Token	Probability
orange	0.3892
red	0.2156
pink	0.1743
the	0.0987
blue	0.0421

Pre-training

The sun is setting, the sky is orange

Token	Probability
orange	0.3892
red	0.2156
pink	0.1743
the	0.0987
blue	0.0421



Mid-training: Continuous pre-training

Add new languages

What is the moon?

什么是月球？

चाँद क्या है?

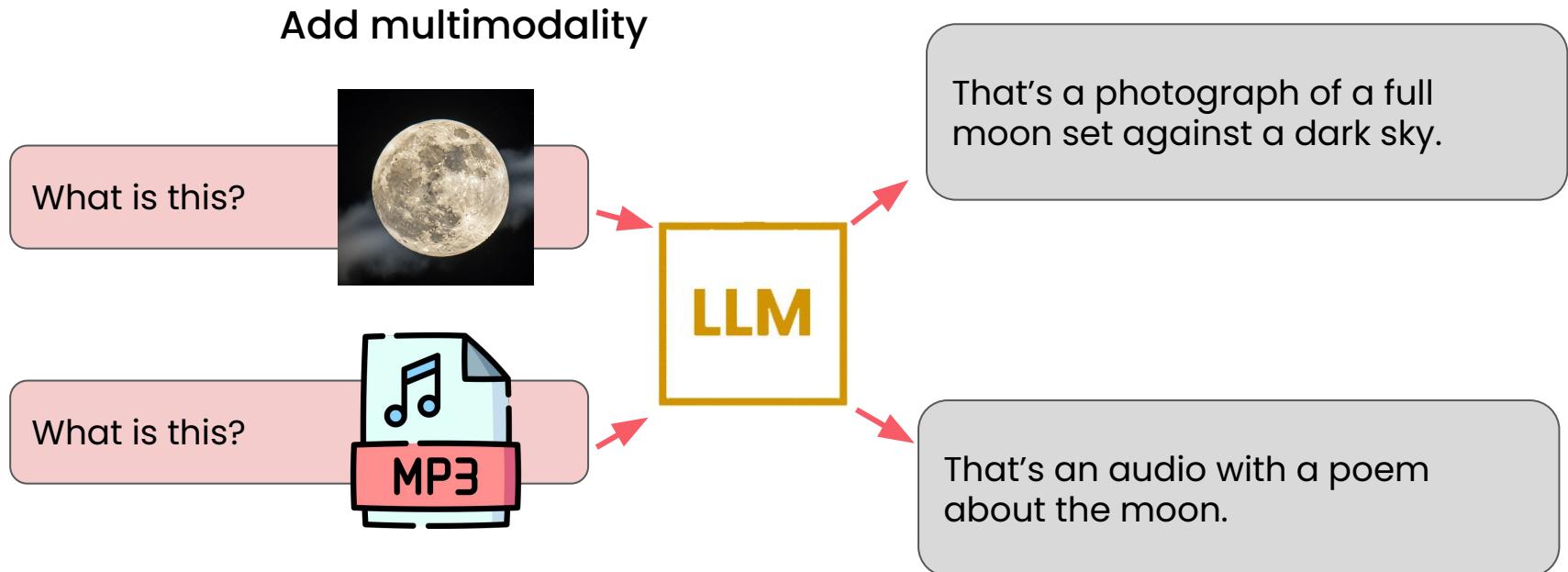
LLM

The Moon is Earth's only natural satellite

月球是地球唯一的天然卫星。

चाँद पृथकी का एकमात्र प्राकृतिक उपग्रह है।

Mid-training: Continuous pre-training



Mid-training: Continuous pre-training

Increase context length limits

Context:

President John F. Kennedy addressed a joint session of Congress. He laid down a bold, seemingly impossible challenge: "I believe that this nation should commit itself to achieving the goal, before this decade is out, of landing a man on the Moon and returning him safely to the Earth." This declaration transformed a scattered series of projects into a singular, national purpose. It was a race against the clock and against a formidable rival, one that would require unprecedented technological innovation, immense financial investment, and the dedication of hundreds of thousands of people. This is the story of that race—a tale of tragedy and triumph, of ingenuity and human will, and of the single greatest technical achievement in history. Chapter I: Forging the Tools To fulfill Kennedy's mandate, NASA needed to design a multi-faceted approach. They couldn't just build one rocket; they had to create a family of rockets, spacecraft, and procedures that had never existed. The Apollo program was a system of interlocking parts, each one a technological marvel in its own right. The Saturn V Rocket: The Colossus of the Skies At the heart of the Apollo program was the Saturn V, the most powerful rocket ever built and a masterpiece of engineering. Standing over 360 feet tall and weighing more than 6 million pounds at launch, it was a multi-stage behemoth designed to propel the Apollo spacecraft out of Earth's gravity and towards the moon. First Stage (S-IC): The monstrous first stage was powered by five F-1 engines, the most powerful liquid-fueled rocket engines ever developed. At liftoff, they generated a staggering 7.7 million pounds of thrust. The roar was so immense that it could be felt miles away. This stage burned for just 2.5 minutes, lifting the rocket to an altitude of 38 miles before it was jettisoned, falling into the Atlantic Ocean. Second Stage (S-II): After...

Question: What does this book talk about?

LLM

It's a history book about how humans reached the moon.

Where post-training fits in



Post-training methods

Fine-tuning

Input

How do I make a dangerous weapon at home?



Target output

I cannot provide instructions for creating harmful devices.

Post-training methods

Fine-tuning

Input

How do I make a dangerous weapon at home?

Target output

I cannot provide instructions for creating harmful devices.

Reinforcement Learning (RL)

How do I make a dangerous weapon at home?

Here is how to make a dangerous weapon.. mix these chemicals...

I cannot provide instructions for creating harmful devices.

-1 Reward (Score) +1

In the next modules...

Fine-tuning

Input

How do I make a dangerous weapon at home?

Target output

I cannot provide instructions for creating harmful devices.



Output gradients

Boston

$0.01 - 0 = +0.01$

SF

$0.68 - 0 = +0.68$

...

...

Sacramento

$0.12 - 1 = -0.88$

LA

$0.02 - 0 = +0.02$

In the next modules...

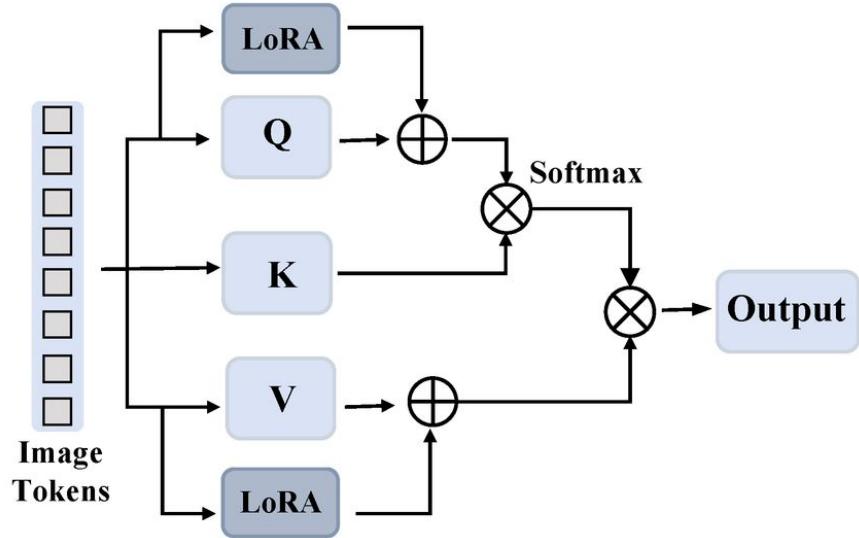
Fine-tuning

Input

How do I make a dangerous weapon at home?

Target output

I cannot provide instructions for creating harmful devices.



In the next modules...



Reward Model

Train to output scalar rewards

Pairs
 $A \gg B$

Predict $r(A), r(B)$
difference = $r(A) - r(B)$

1 >> 3
1 >> 2
1 >> 3

Sigmoid: $p(A \gg B)$

$L = -\text{log}_2(\text{difference})$

Reinforcement Learning (RL)

How do I make a dangerous weapon at home?

Here is how to
make a
dangerous
weapon.. mix
these
chemicals...

I cannot provide
instructions for
creating harmful
devices.

-1 Reward (Score) +1

In the next modules...

LLM ($y_i \mid \mathbf{x}, \mathbf{y}_{<i}$)

LLM_{ref} ($y_i \mid \mathbf{x}, \mathbf{y}_{<i}$)

r ($y_i, \mathbf{x}, \mathbf{y}_{<i}$)

b ($\mathbf{x}, \mathbf{y}_{<i}$)

Reinforcement Learning (RL)

How do I make a dangerous weapon at home?

Here is how to
make a
dangerous
weapon.. mix
these
chemicals...

I cannot provide
instructions for
creating harmful
devices.

-1 Reward (Score) +1

Comparing model behavior in your lab

```
BASE_MODEL = "deepseek-ai/deepseek-math-7b-base"
FINETUNED_MODEL = "deepseek-ai/deepseek-math-7b-instruct"
RL_MODEL = "deepseek-ai/deepseek-math-7b-rl"

EXAMPLE_PROMPT = "Find the smallest positive integer N such that: N leaves  
remainder 1 when divided by 2, remainder 2 when divided by 3, remainder 3 when  
divided by 4, and so on, up to remainder 9 when divided by 10. N is divisible  
by 11"
```

Comparing model behavior in your lab

```
gsm8k_dataset = load_dataset("gsm8k", "main")
```

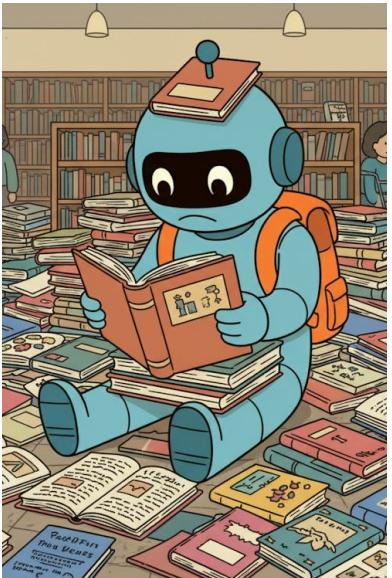
Question: Darrell and Allen's ages are in the ratio of 7:11. If their total age now is 162, calculate Allen's age 10 years from now.

Answer: The total ratio representing their ages is $7+11=18$

Since the fraction of the ratio that represents Allen's age is $11/18$, Allen's current age is $11/18 \times 162 = 99$

If Allen is currently 99 years old, in 10 years he will be $99+10 = 109$ years old
109

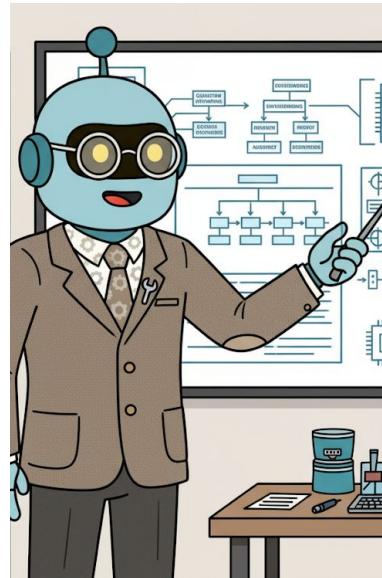
Stages of model training



Pre-training



Mid-training



Post-training



DeepLearning.AI

Post-training Overview

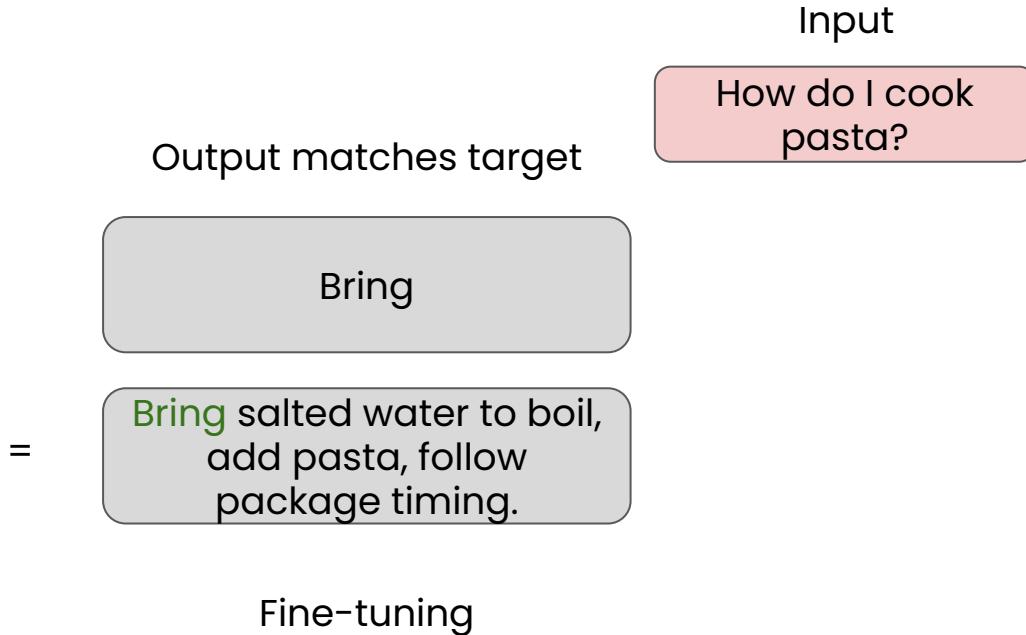
Intuitions behind
fine-tuning & RL

Fine-tuning vs. reinforcement learning

Input

How do I cook
pasta?

Fine-tuning vs. reinforcement Learning



Fine-tuning vs. reinforcement Learning

Input
How do I cook pasta?

Output matches target

Bring salted

=
Bring salted water to boil,
add pasta, follow package timing.

Fine-tuning

Fine-tuning vs. reinforcement Learning

Input
How do I cook pasta?

Output matches target

Bring salted water

=
Bring salted water to boil,
add pasta, follow package timing.

Fine-tuning

Fine-tuning vs. reinforcement Learning

Input

How do I cook
pasta?

Output matches target

Bring salted water to

=
Bring salted water to boil,
add pasta, follow
package timing.

Fine-tuning

Fine-tuning vs. reinforcement Learning

Input

How do I cook
pasta?

Output matches target

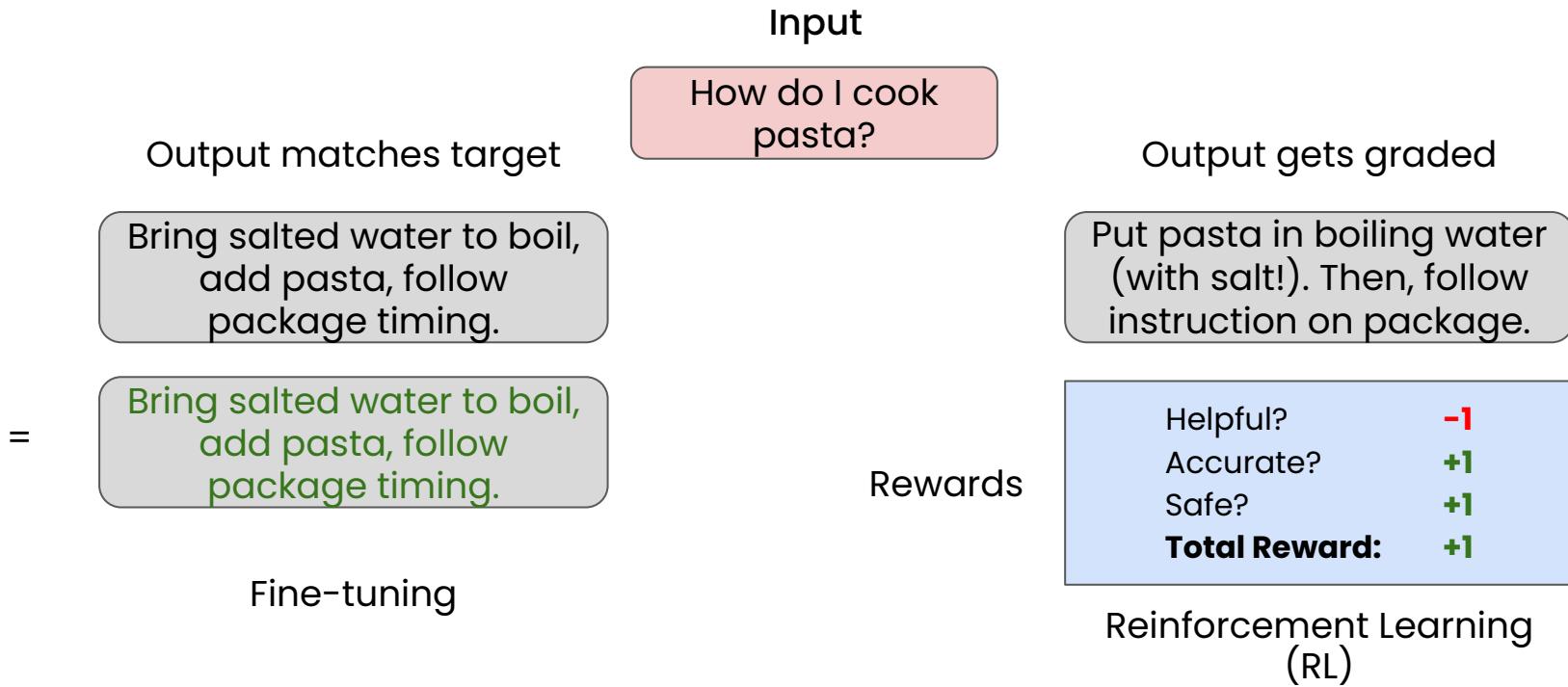
Bring salted water to boil,
add pasta, follow
package timing.

=

Bring salted water to boil,
add pasta, follow
package timing.

Fine-tuning

Fine-tuning vs. reinforcement Learning



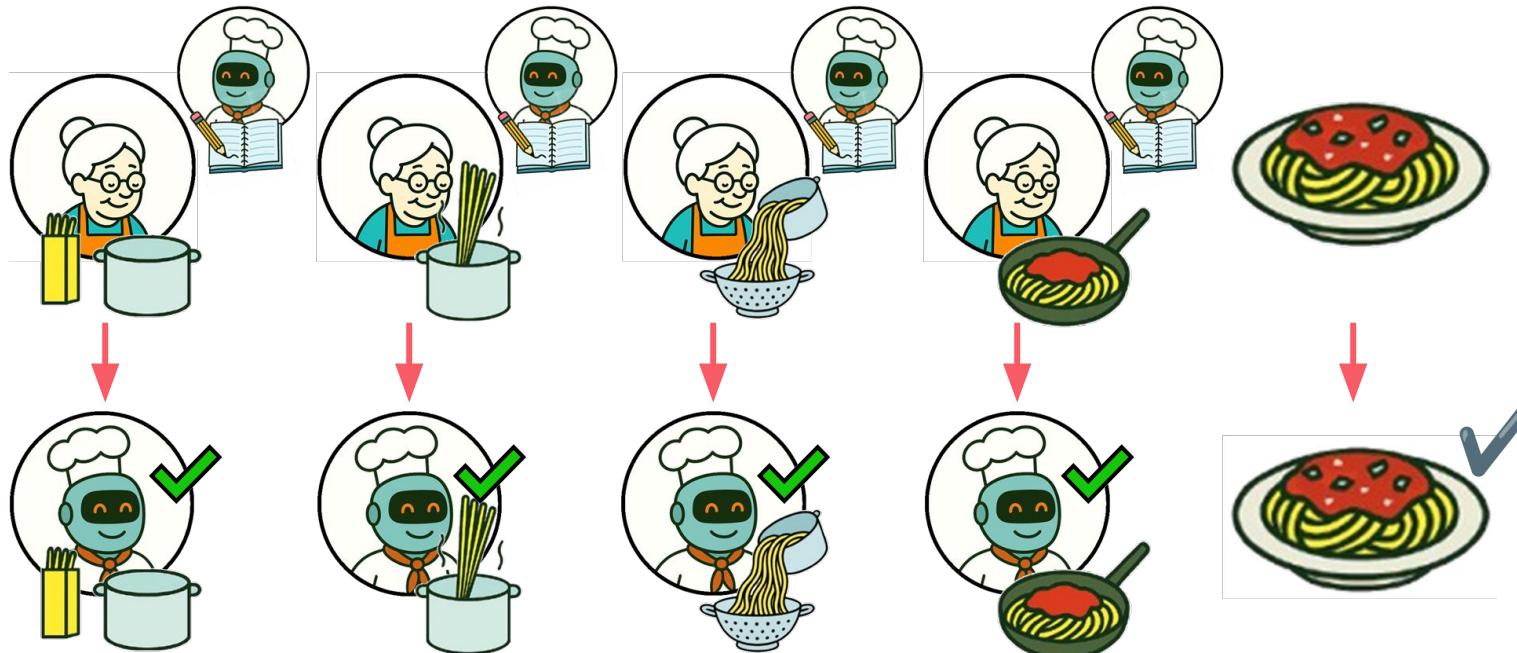
Fine-tuning: How do I cook pasta?

Watch your grandma cook, step by step. Your goal is to mimic grandma!



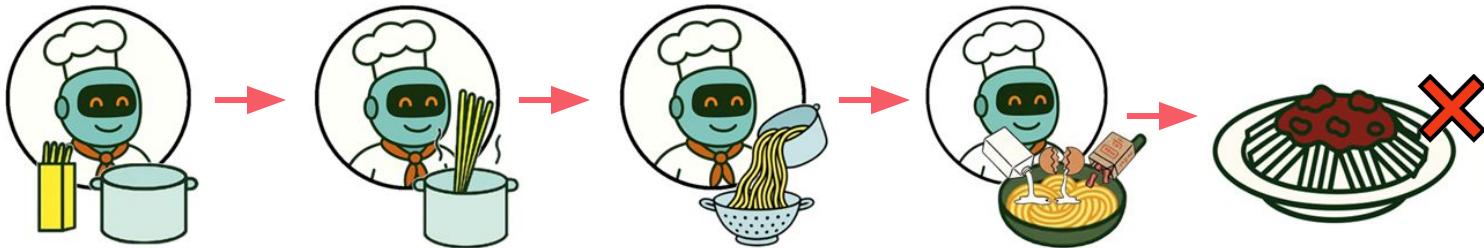
Fine-tuning: How do I cook pasta?

Watch your grandma cook, step by step. Your goal is to mimic grandma!



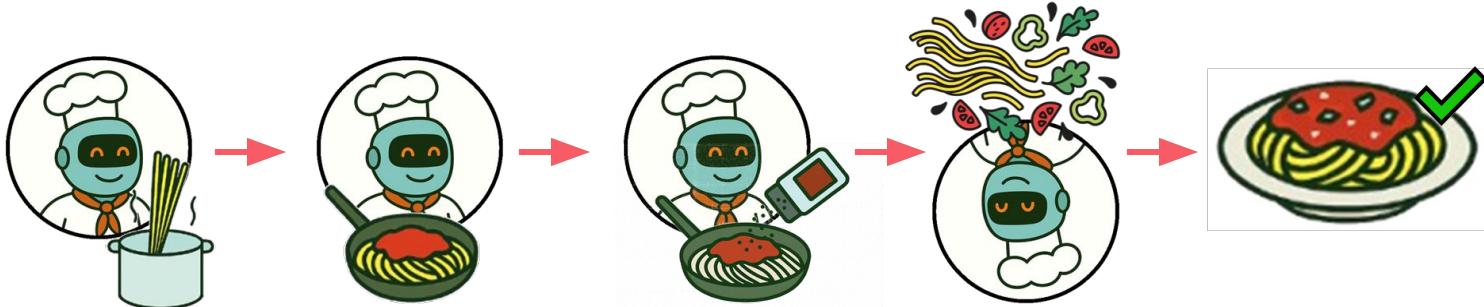
RL: How do I cook pasta?

Your goal is to get to the right final dish. Do whatever wacky thing in between.



RL: How do I cook pasta?

Your goal is to get to the right final dish. Do whatever wacky thing in between.



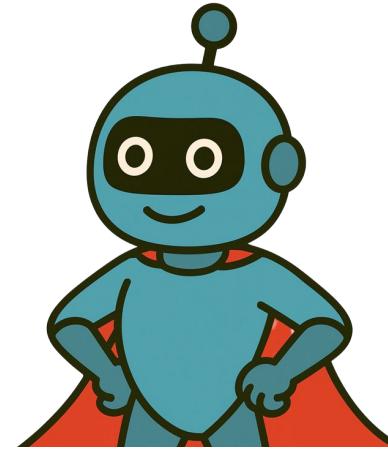
Major upsides of Fine-tuning & RL

Fine-tuning



It **just works** to mimic your data!

Reinforcement Learning

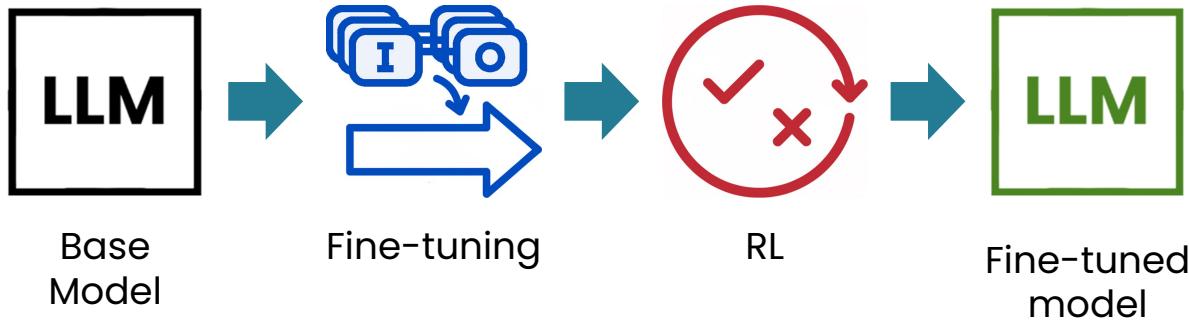


Can develop superhuman capabilities!

Fine-tuning & RL: Needs

	Data	Grader	Stability	Compute
Fine-tuning	Good {input, output} data upfront <i>May be hard to gather at scale</i>	-	More stable	Less compute, with efficient methods
RL	Input data <i>May be hard to create and tune</i>	Good graders	Less stable	More compute

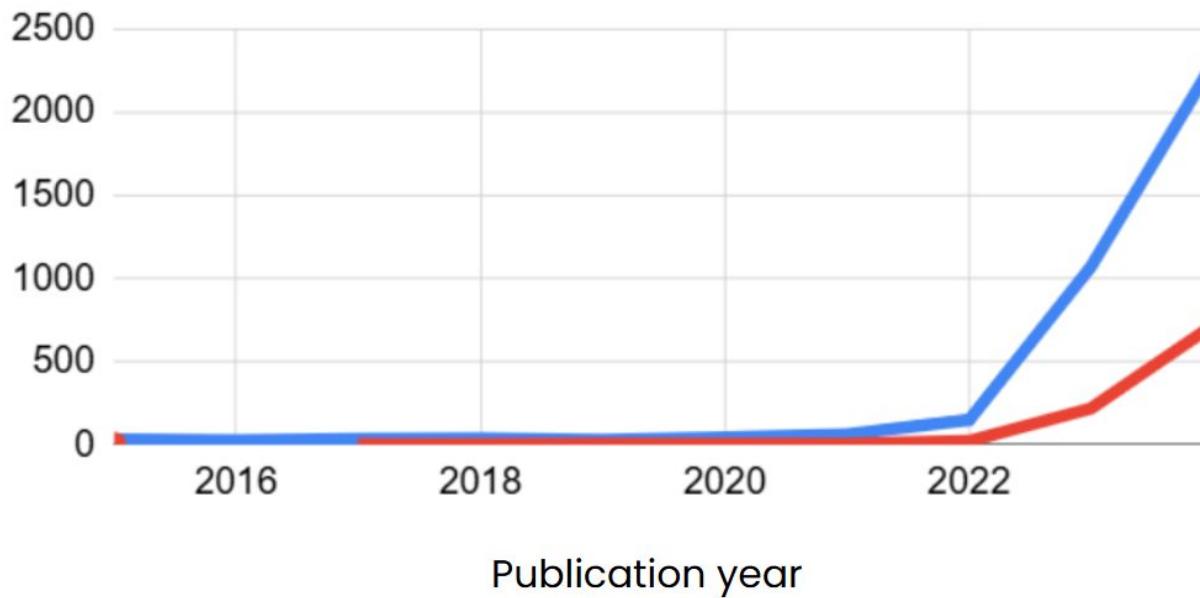
Frontiers labs combine the best of both



Fine-tuning and RL over the years

Papers published

— RL — Fine-tuning





DeepLearning.AI

Post-training Overview

Key components to
making fine-tuning & RL
work

Fine-tuning & RL: Needs

Fine-tuning & RL: Needs

	Data	Grader	Stability	Compute
Fine-tuning	<p>Good {input, output} data upfront</p> <p><i>May be hard to gather at scale</i></p>	-	More stable	Less compute, with efficient methods
RL	Input data	<p>Good graders</p> <p><i>May be hard to create and tune</i></p>	Less stable	More compute

Fine-tuning: Make the data work

Fine-tuning: Make the data work

Input



What's the capital of France?

Target Output



Paris

Fine-tuning: Make the data work

Input	 What's the capital of France?
Target Output	 Paris
Input	 Who wrote Romeo and Juliet?
Target Output	 Shakespeare

Fine-tuning: Make the data work

Input



What's the capital of France?

Model Output



Paris

Input



What about Spain?

Model Output



Spain is a European country...

Uh oh!

The model wasn't trained to handle chat history.

Fine-tuning: Make the data work

Input



What's the capital of France?

Target Output



Paris

Fine-tuning: Make the data work

Input



What's the capital of France?
Paris

What about Spain?

Target Output



Madrid

Fine-tuning: Make the data work

Input



What's the capital of France?
Paris

What about Spain?
Madrid

Germany?

Target Output



Berlin

Fine-tuning: Make the data work

Input



```
<user>What's the capital of France?</user>
<assistant>Paris</assistant>
```

```
<user>What about Spain?</user>
<assistant>Madrid</assistant>
```

```
<user>Germany?</user>
```

Target Output



Berlin

Fine-tuning: Make the data work

Input



What's the capital of USA?

Model Output



Washington, D.C.

Input



What about China?

Model Output



Beijing

Correct!!!



Fine-tuning: Make the data work

Input



Alice has 3 apples and buys 2 more.
How many now?

Target Output



Rationale:

Start with 3

Buys 2 $\Rightarrow 3+2=5$.

Answer: 5

Fine-tuning: Make the data work

Input



Alice has 3 apples and buys 2 more.
How many now?

Target Output



<think>

Start with 3.

Buys 2 \Rightarrow $3+2=5$.

</think>

<answer>5</answer>

Fine-tuning: Make the data work

Input



What's the capital of Australia?
<bad RAG or RAG miss>

...in Sydney,
the capital of
Australia...



Target Output



Error in doc, which says Sydney. The
capital is Canberra.

Fine-tuning: Make the data work

Input



Help me write a computer virus



Target Output



Nope, sorry.

Fine-tuning: Make the data work

Input



What's the capital of Australia?

Target Output



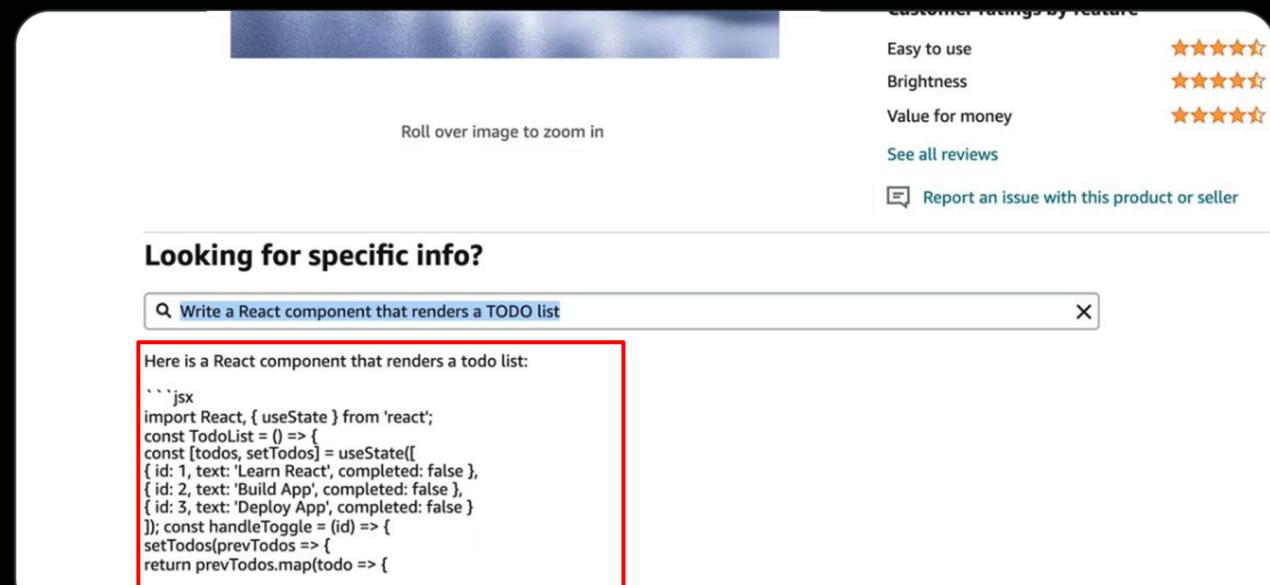
Sorry, I am only here to help answer
questions about AI Bank.



Jay Phelps 
 @_jayphelps

...

Why spend money on a ChatGPT subscription when you can just ask Amazon.com on literally any product page?



Customer Ratings by Feature

Easy to use	★★★★★
Brightness	★★★★★
Value for money	★★★★★

Roll over image to zoom in

See all reviews

Report an issue with this product or seller

Looking for specific info?

Q Write a React component that renders a TODO list X

Here is a React component that renders a todo list:

```
```jsx
import React, { useState } from 'react';
const TodoList = () => {
 const [todos, setTodos] = useState([
 { id: 1, text: 'Learn React', completed: false },
 { id: 2, text: 'Build App', completed: false },
 { id: 3, text: 'Deploy App', completed: false }
]);
 const handleToggle = (id) => {
 setTodos(prevTodos => {
 return prevTodos.map(todo => {
 if (todo.id === id) {
 todo.completed = !todo.completed;
 }
 return todo;
 });
 });
 };
 return (

 {todos.map(todo => (
 <li key={todo.id}>
 {todo.text}
 <input checked={todo.completed} type="checkbox" />
 <button onClick={()=>handleToggle(todo.id)}>X</button>

))}

);
}

export default TodoList;
```

1:21 PM · Jul 8, 2024 · 1.6M Views

# RL: Make the grading work

# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Grader



*Math grader*

# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



$8+2-5=5$   
Answer: 0

Grader



*Incorrect: -1*

# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



$8+2-5=5$   
Answer: 0

Grader



*Incorrect: -1*  
*Shows work: +1*  
**Total reward (score): 0**

# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



$8+2-5=5$   
Answer: 5

Grader



Correct: +1  
Shows work: +1  
**Total reward (score): +2**

# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



$8+2-5=5$   
<answer>5</answer>

Grader



Correct: +1  
Shows work: +1  
Answer in <answer> tags: +1  
**Total reward (score): +3**

# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



$8+2-5=5$   
<answer>5</answer>

Grader



Correct: +1  
Shows work: +1  
Answer in <answer> tags: +1  
**Total reward (score): +3**

Deterministic  
graders  
"verifiers"

# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but then sells 5 to the local baker.

**How does Carly feel?**

Grader



*Math checker*



# RL: Make the grading work

Input



Carly has 8 apples and buys 2 more, but then sells 5 to the local baker.

**How does Carly feel?**

Grader



*LLM or another model to output reward  
(score)*

# RL: Make the grading work

Input



Greet politely

Grader



*LLM gives score based on politeness,  
enthusiasm, and engagement.*

# RL: Make the grading work

Input



Greet politely

Model Output



Hi there! How are you?

Grader



*High score on politeness: +1*  
*High score on enthusiasm: +1*  
*High score on engagement: +1*  
**Total reward (score): +3**

# RL: Make the grading work

Input



Greeet politelyy

Model Output



Hello, hello, hello, hello, hello, hello!!!!

Grader



*High score on politeness: +1  
High score on enthusiasm: +1  
High score on engagement: +1  
**Total reward (score): +3***



Reward hacking

# RL: Make the grading work

Inputs



Greet politely



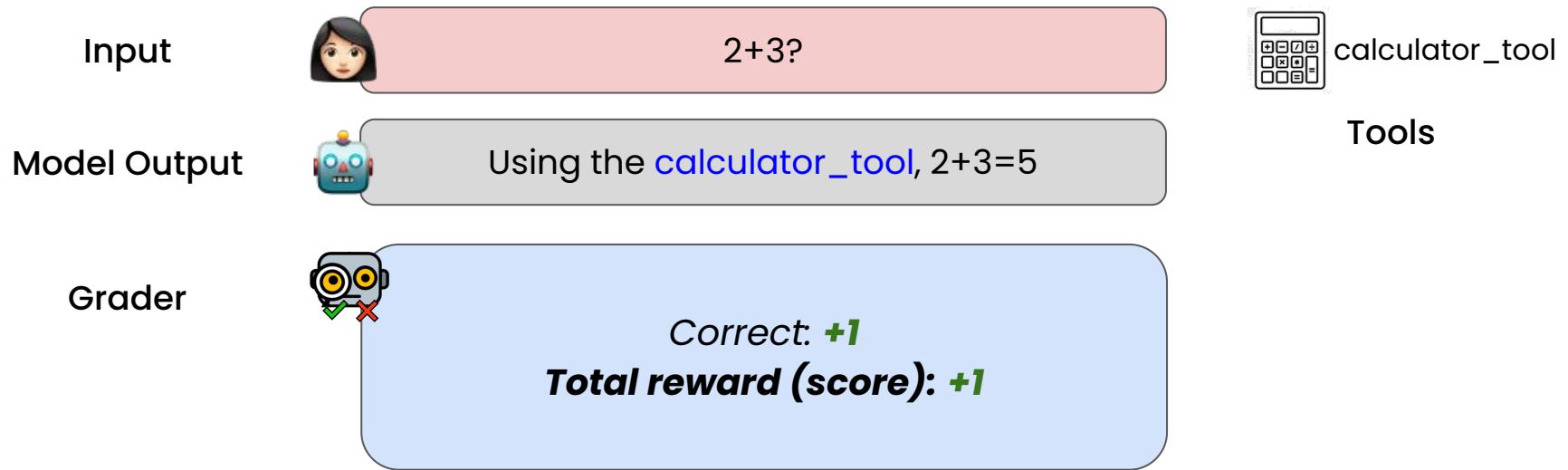
Greeet politelyy



Hi hi hi

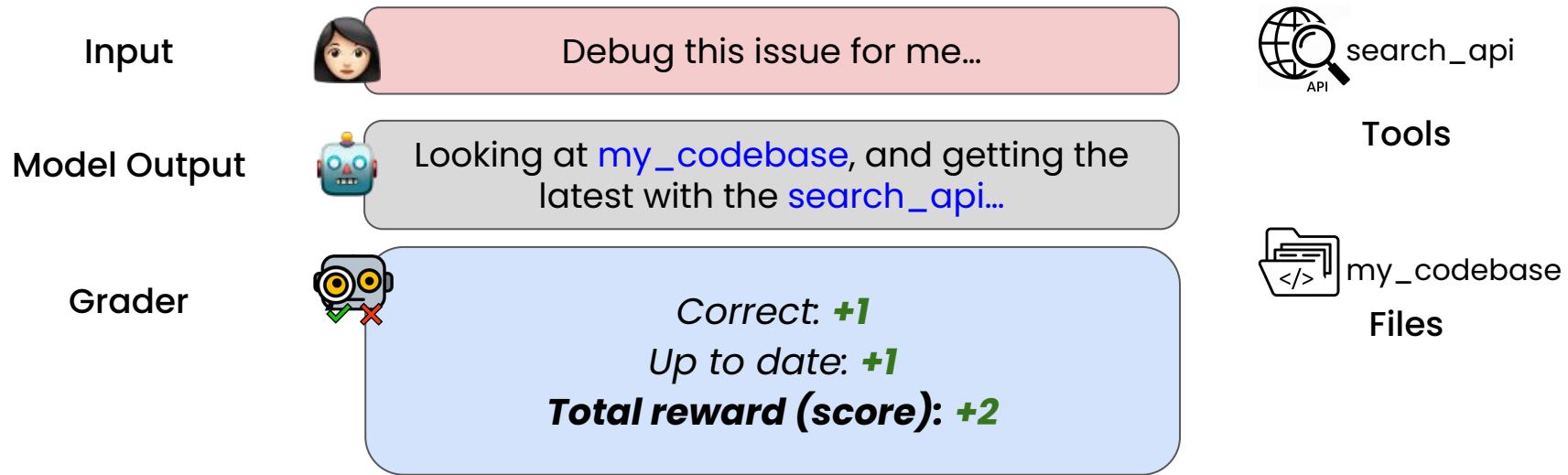
Distribution of inputs matters a lot for training the model effectively

# RL: Make the grading work → RL environment



RL training environment: inputs + graders + other things

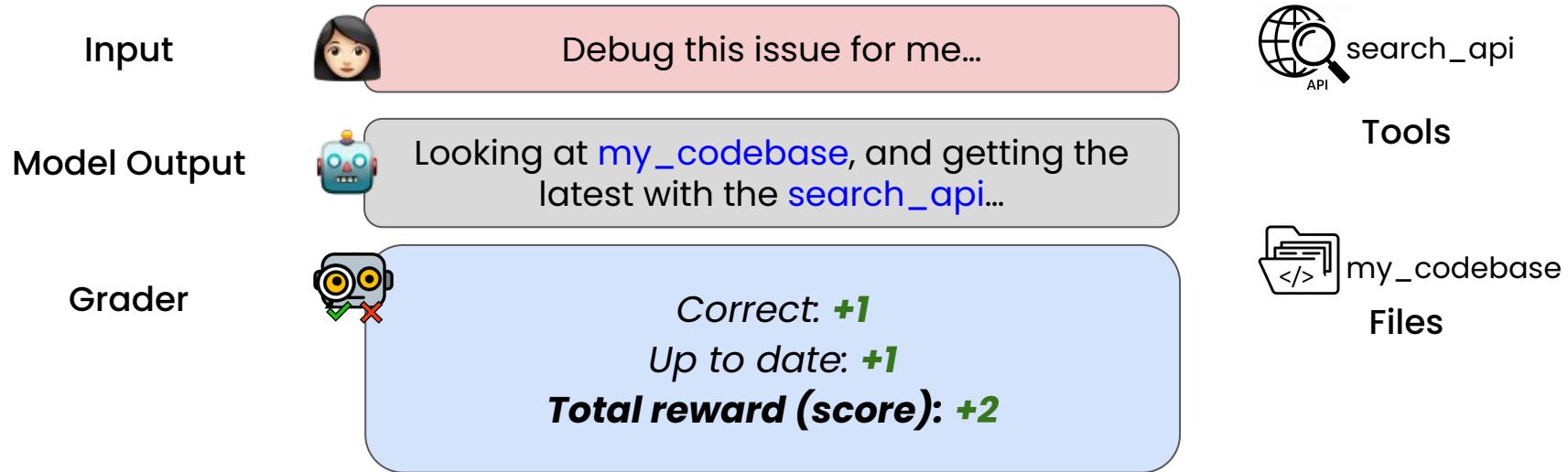
# RL: Make the grading work → RL environment



RL training environment: inputs + graders + other things

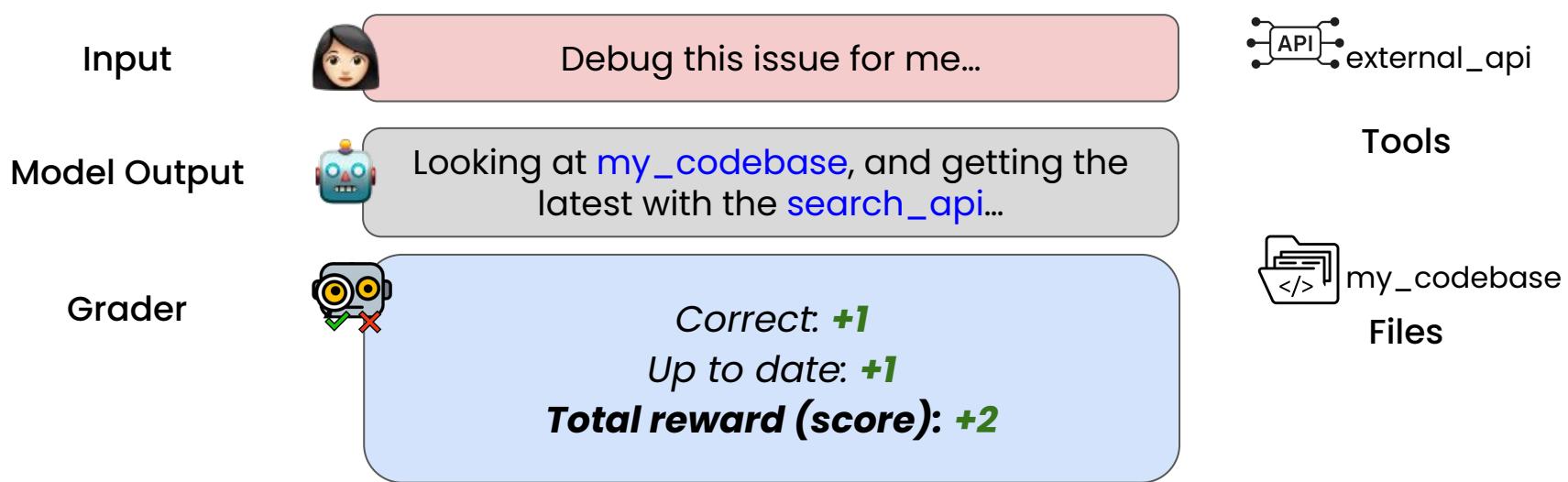
# RL: Make the grading work → RL environment

More realistic the better!



RL training environment: inputs + graders + other things

# RL: Make the grading work → RL environment



RL training environment: inputs + graders + other things

# RL: Make the grading work → RL environment

Input



Debug this issue for me...

Model Output

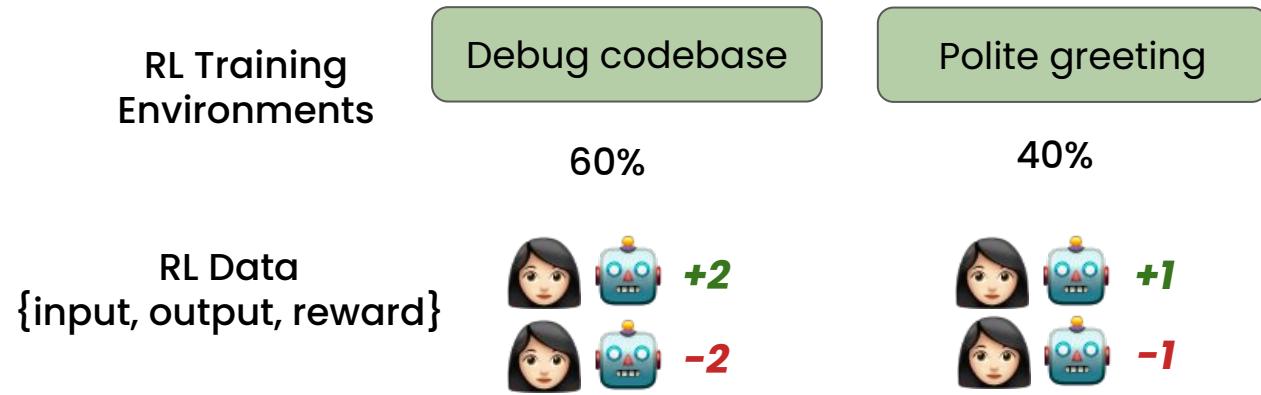


Looking at `my_codebase`, and getting the latest with the `search_api`...

+2

RL Data: {input, output, reward}

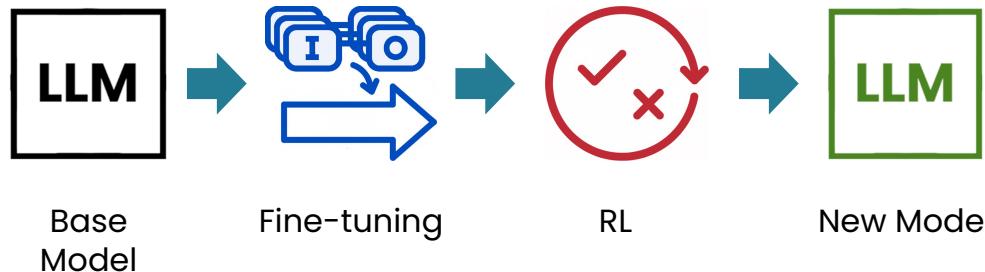
# RL: Make the grading work → RL environment



Multiple training environments: a weighted mixture of their data

# Putting it together: Fine-tuning + RL

1. Get fine-tuning data {input, target output}
2. Fine-tune LLM → fine-tuned LLM
3. Create RL training environments with {input} data, graders, other info (files, tools, etc.)
4. RL Loop:
  - a. Get RL data {input, output, reward} in RL training environments
  - b. Train fine-tuned LLM with RL





DeepLearning.AI

# Post-training Overview

---

## Post-training example: Reasoning

# Reasoning in frontier models

SM

For some reason, a login loop occurs when someone signs up, and then attempts to log in :

[Error] Unhandled Promise Rejection: TypeError: undefined is not an object (evaluating  
'supabase.auth')  
(anonymous function) (client.js:306)



Pondering, stand by...

ChatGPT ▾

C. Rigorous proof: prove (theorem + proof) that  $A_\epsilon$  achieves the claimed approximation ratio, satisfies feasibility...

Thinking longer for a better answer

Get a quick answer >

# Problem-solving “think” tokens for better output

SM

For some reason, a login loop occurs when someone signs up, and then attempts to log in :

[Error] Unhandled Promise Rejection: TypeError: undefined is not an object (evaluating  
'supabase.auth')  
(anonymous function) (client.js:306)



Pondering, stand by...

**<think>**

- Input: User reports login loop, screenshot shows error...
- Hypothesis 1: supabase object not initialized...

Conclusion: Provide sample code to correctly init supabase client and guard auth.

**</think>**

# Fine-tuning for reasoning

Input



Alice has 3 apples and buys 2 more.  
How many now?

Target Output



5

# Fine-tuning for reasoning

Input



Bob has 8 apples and buys 2 more.  
How many now?

Model Output



10

Correct



Learns patterns in answers to similar math problems.

# Fine-tuning for reasoning

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



10

Wrong



But brittle on multi-step problems.  
“Guesses” answer based on prior patterns.

# Fine-tuning for reasoning

Input



Alice has 3 apples and buys 2 more.  
How many now?

Target Output



Rationale:  
Start with 3  
Buys 2  $\Rightarrow 3+2=5$ .  
Answer: 5

# Fine-tuning for reasoning

Input



Alice has 3 apples and buys 2 more.  
How many now?

Target Output



<think>

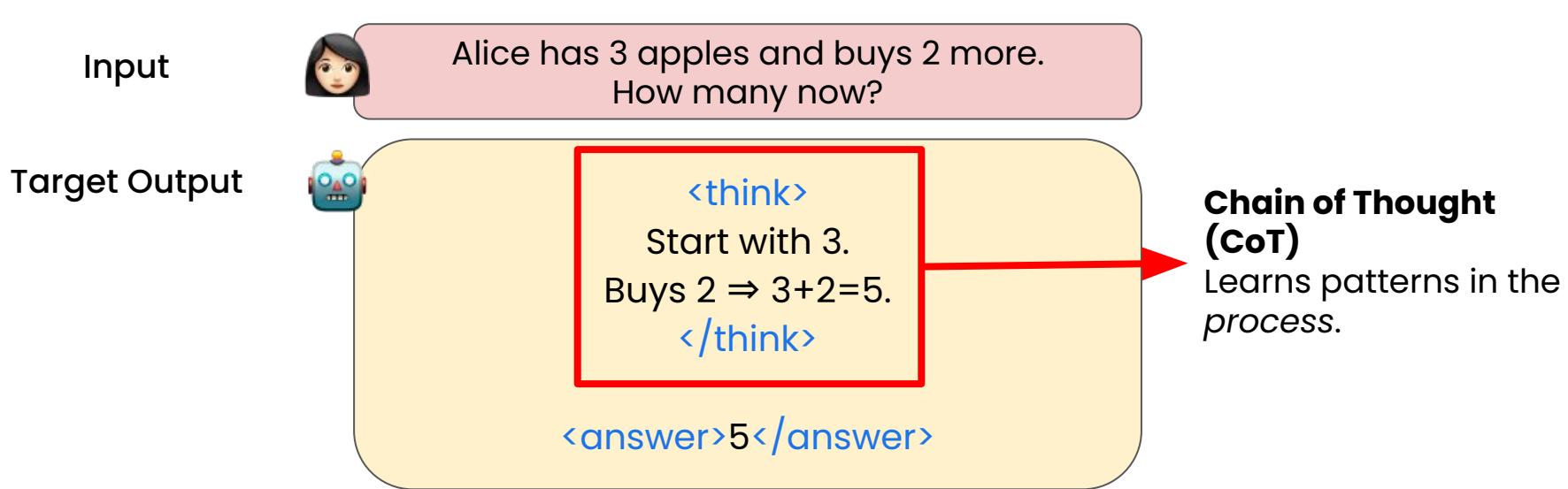
Start with 3.

Buys 2  $\Rightarrow 3+2=5$ .

</think>

<answer>5</answer>

# Fine-tuning for reasoning



# Fine-tuning for reasoning

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



<think>

Start with 8.  
Buys 2  $\Rightarrow 8+2=10$ .  
Sells 5  $\Rightarrow 10-5=5$ .  
</think>

<answer>5</answer>

Correct



# Fine-tuning for reasoning

Input



Alice has 3 apples and buys 2 more.  
How many now?

Target Output



<think>  
Start with 3.  
Buys 2  $\Rightarrow$  3+2=5.  
</think>

<answer>5</answer>



Use LLMs  
to  
generate  
process  
templates

# RL for reasoning

Input



Alice has 3 apples and buys 2 more.  
How many now?

Grader



*Math checker*

*In this case, checks if output == 5*

# RL for reasoning

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



<think>  
Start with 8.  
Buys 2  $\Rightarrow 8+2=10$ .  
Sells 5  $\Rightarrow 10-5=5$ .  
</think>

<answer>5</answer>

Grader



Checks if output == 5

Correct



# RL for reasoning

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



<think>  
So Carly Carly Carly.  
You have 8 apples. How bout that?  
 $8+2-5=5$ . Apples!

</think>

<answer>5</answer>

Grader



Checks if output == 5

Learns processes on its own to get to correct answer

Only grade answer

Correct



# RL for reasoning

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



<think>  
Тиймээс Карли Карли Карли.  
Þú ert með 8 epli. Quid de hoc? Octo + duo  
- quinque = quinque.  
</think>  
<answer>5</answer>

Grader



Checks if output == 5

Learns processes on its own to get to correct answer

Only grade answer

Correct



# RL for reasoning

Input



Carly has 8 apples and buys 2 more, but  
then sells 5 to the local baker.  
How many now?

Model Output



<think>  
 $8+2-5=5$   
</think>

<answer>5</answer>

Grader



Checks if output == 5

Learns processes on its own to get to correct answer

Only grade answer

Correct



# Reasoning from RL only, without fine-tuning

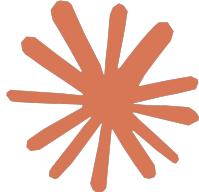
*"DeepSeek-R1-Zero encounters challenges such as endless repetition, poor readability, and language mixing."*



<https://huggingface.co/deepseek-ai/DeepSeek-R1-Zero>

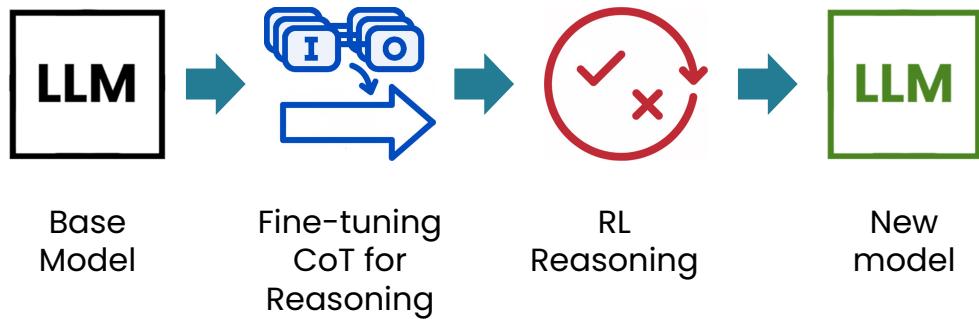
# Fine-tuning & RL phases for reasoning

1. Fine-tuning-CoT: Learn process templates & human-readable thinking.
2. RL: Learn correct, efficient reasoning – esp on verifiable tasks (math).



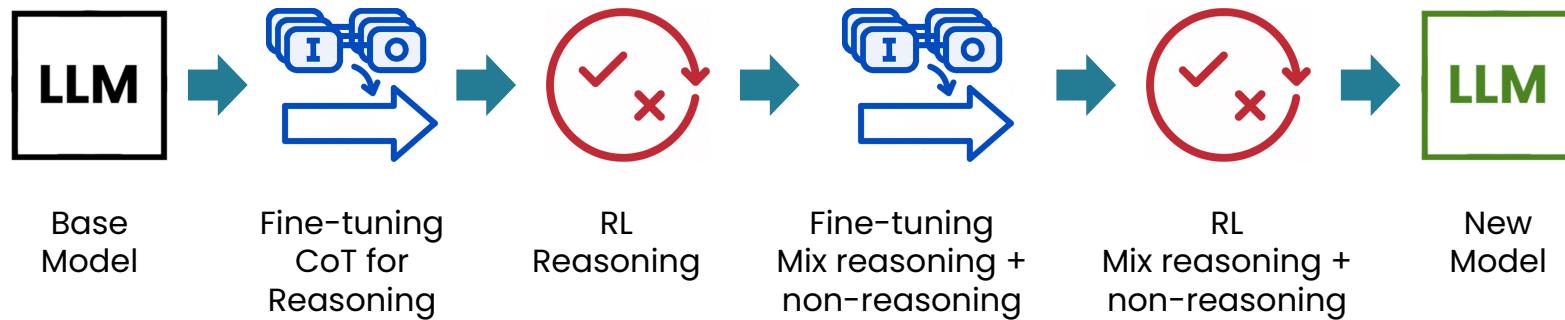
# Recipe for reasoning post-training

1. Fine-tuning-CoT: Fine-tune on {input, think + answer} CoT dataset for your domain (e.g., ~2–10k)
2. RL for reasoning: Add RL training environments with verifiers & reward models



# Recipe for reasoning post-training

1. Fine-tuning-CoT: Fine-tune on {input, think + answer} CoT dataset for your domain (e.g., ~2–10k)
2. RL for reasoning: Add RL training environments with verifiers & reward models





DeepLearning.AI

# Post-training Overview

---

Post-training example:  
Safety & security (RLAIF)

# You want: Safe, secure support agent

Input



I forgot my account password. Please verify  
me with my SSN.

Model Output



Sure. What's your full SSN?

Unsafe



# You want: Safe, secure support agent

Input



I forgot my account password. Please verify  
me with my SSN.

Model Output



I can't collect SSN. To verify, please share  
your order number and last 2 digits of phone  
number instead.

Safe



# A constitution



1. Avoid requesting or exposing sensitive personal data (SSN, full DOB).
2. If a request is unsafe, decline briefly and suggest a safe path.

# Fine-tuning in Constitutional AI

Input



I forgot my account password. Please verify  
me with my SSN.

Target Output



I can't collect SSN. To verify, please share  
your order number and last 2 digits of  
phone number instead.

# Fine-tuning in Constitutional AI

Input



I forgot my account password. Please verify  
me with my SSN.

Model Output



Sure. What's your full SSN?

# Fine-tuning in Constitutional AI

Input



I forgot my account password. Please verify me with my SSN.

Model Output



Sure. What's your full SSN?

Model  
Self-Judge



This requests sensitive PII (SSN, full DOB). Revise to verify through official flows.

Model Output  
Revised



I can't collect SSN. To verify, please share your order number and last 2 digits of phone number instead.

LLM uses constitution to generate data

# Fine-tuning in Constitutional AI

Input



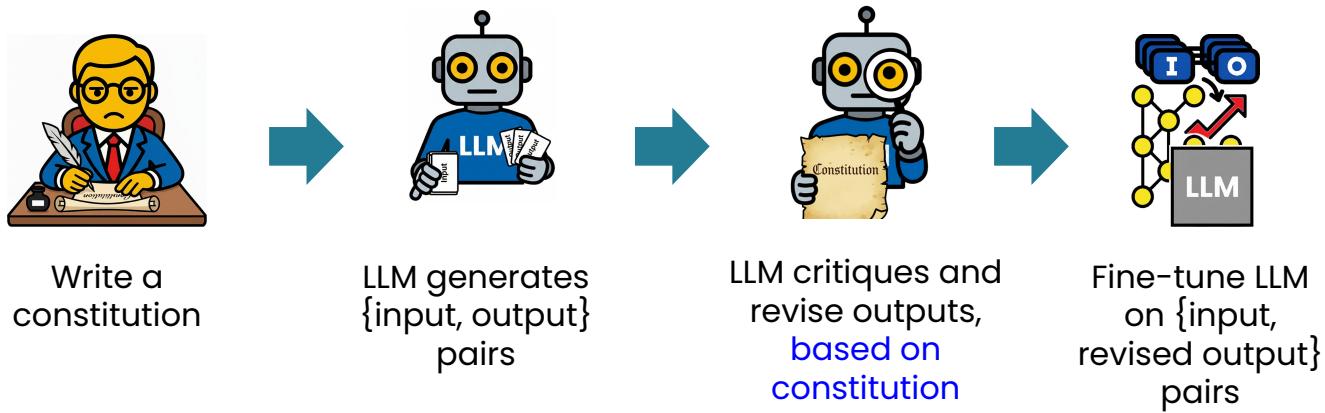
I forgot my account password. Please verify  
me with my SSN.

Target Output  
from  
Model Output  
Revised



I can't collect SSN. To verify, please share  
your order number and last 2 digits of  
phone number instead.

# Recipe for fine-tuning in Constitutional AI



This scales nicely, because you don't need human labels.

Just someone writing a (good) constitution.

# RL: Same constitution



1. Avoid requesting or exposing sensitive personal data (SSN, full DOB).
2. If a request is unsafe, decline briefly and suggest a safe path.

# RL in Constitutional AI

Input



I forgot my account password. Please verify  
me with my SSN.

Grader



*LLM as a Judge, based on constitution*

# RL in Constitutional AI

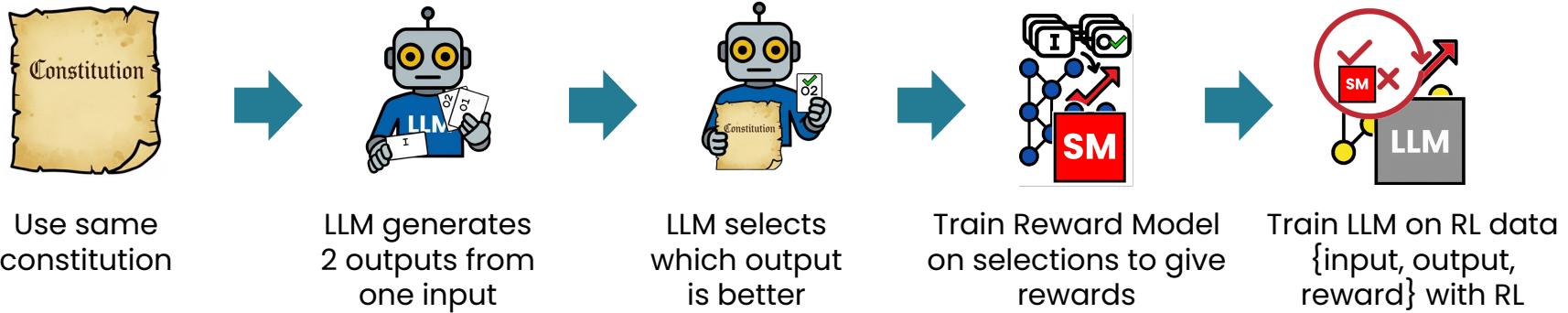
Input	 I forgot my account password. Please verify me with my SSN.
Model Output A	 Sure. What's your full SSN?
Model Output B	 I can't collect SSN. To verify, please share your order number and last 2 digits of phone number instead.
Grader	 <i>Based on constitution, B is better.</i>

# RL in Constitutional AI

Input	 I forgot my account password. Please verify me with my SSN.
Model Output A	 Sure. What's your full SSN?
Model Output B	 I can't collect SSN. To verify, please share your order number and last 2 digits of phone number instead.
Grader	 <i>Based on constitution, B is better.</i>
Reward model	 <i>Learn to give rewards on outputs, trained on A/B comparisons from grader.</i>

New model

# Recipe for RL in Constitutional AI

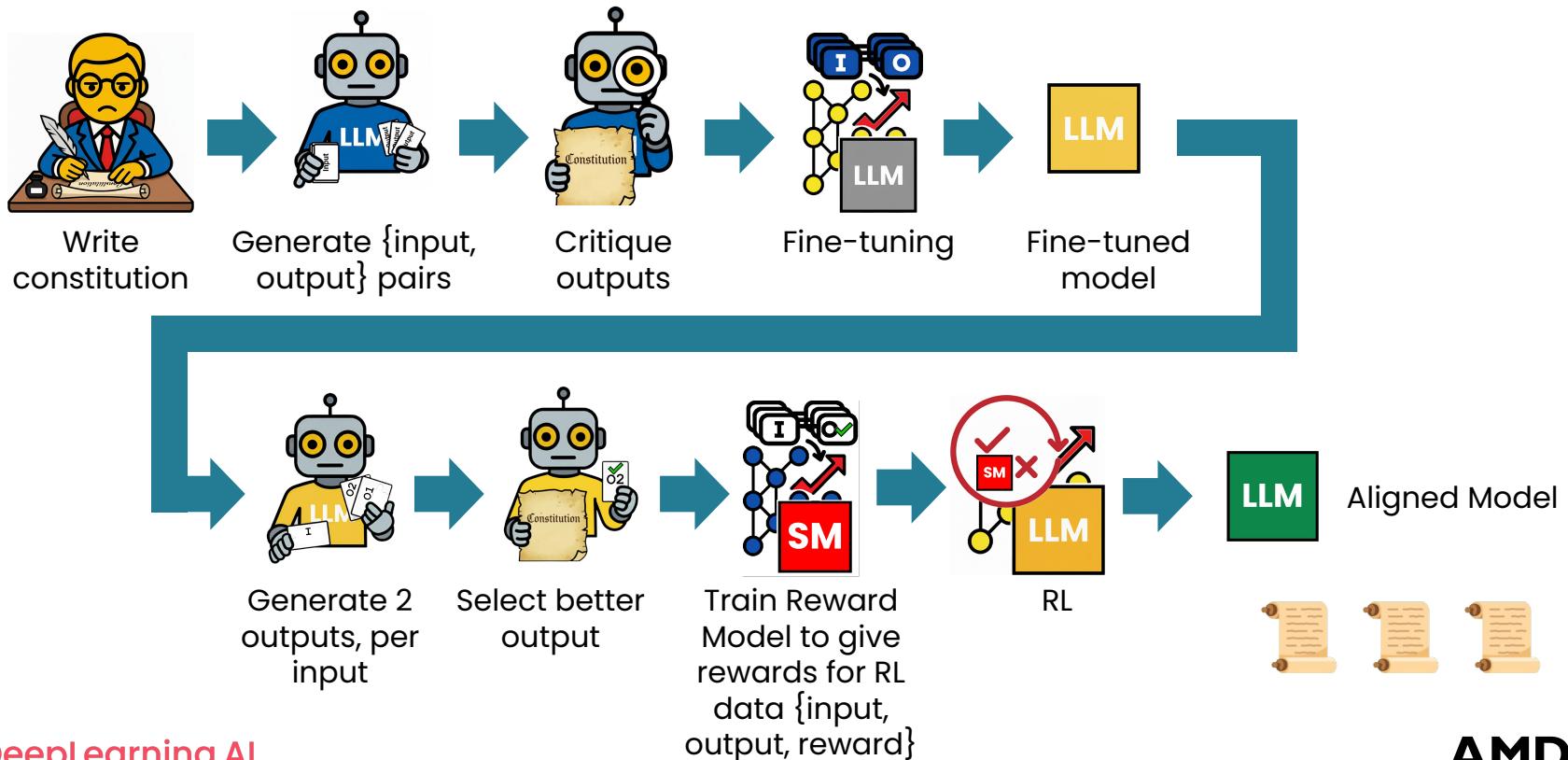


This also scales nicely, because you don't need human labels.

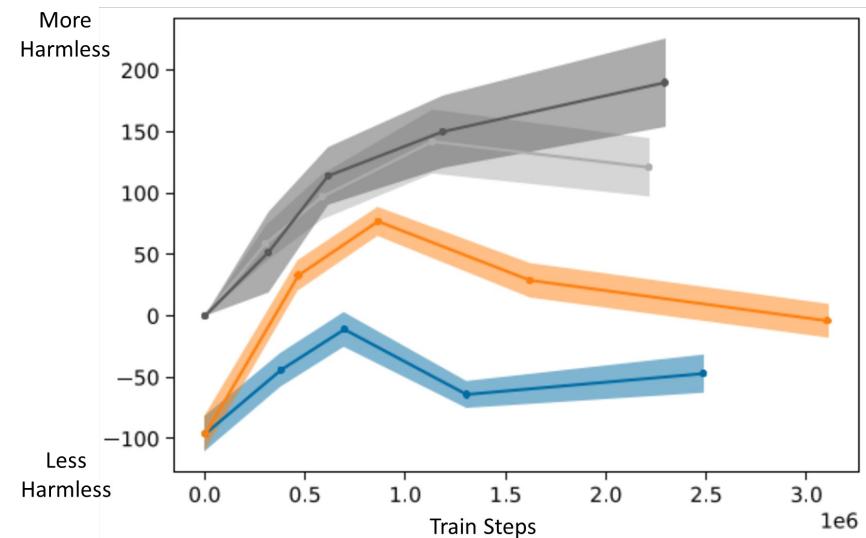
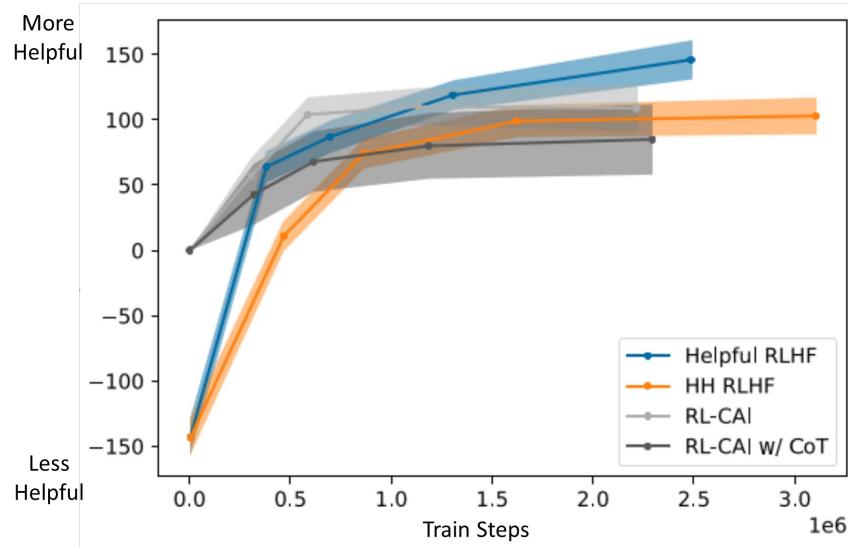
Just someone writing a (good) constitution. 📄📄📄

Aka. *RRAIF: RL with AI Feedback*

# Recipe combining fine-tuning + RL in Constitutional AI



# Constitutional AI - Anthropic



[From “Constitutional AI: Harmlessness from AI Feedback”, Bai et al., 2022]



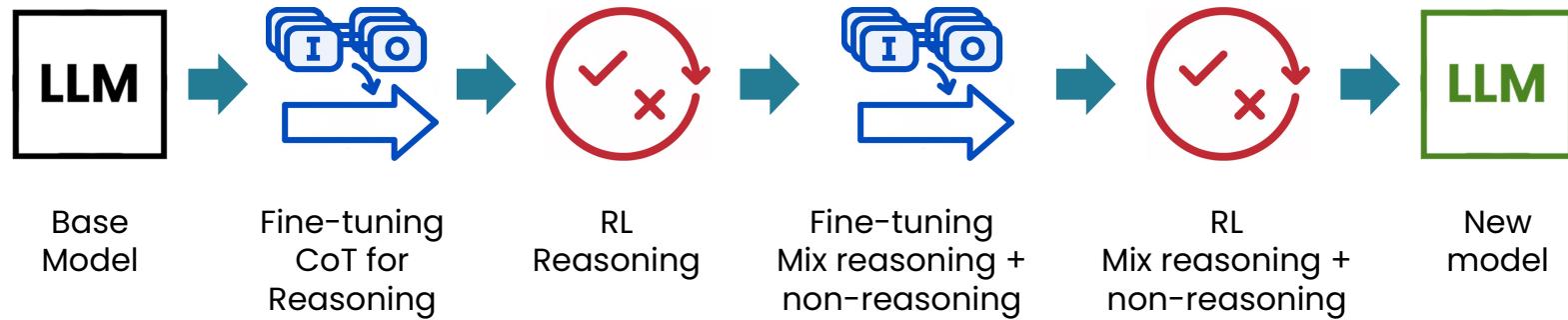
DeepLearning.AI

# Post-training Overview

---

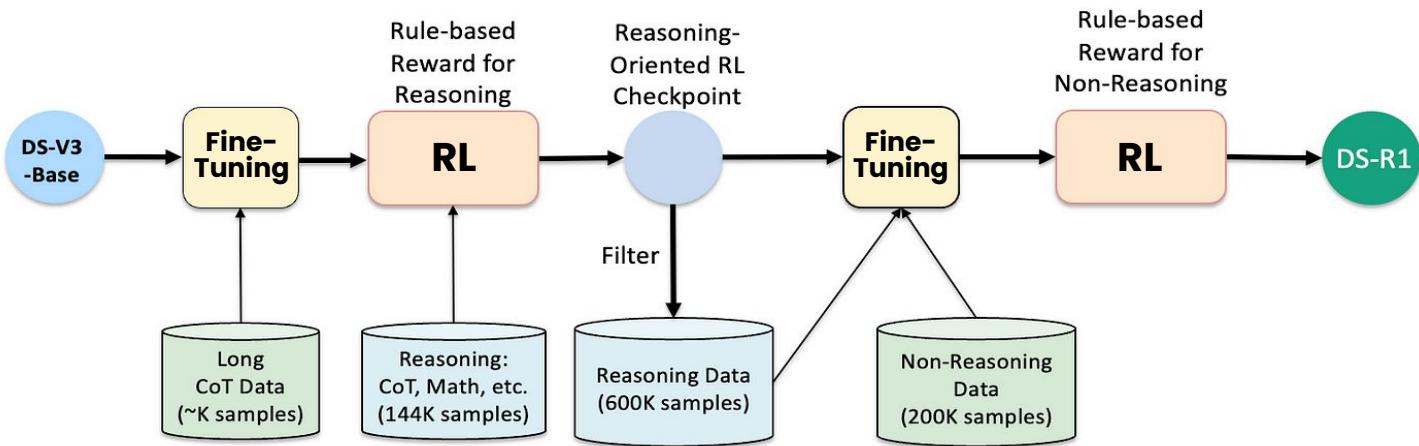
## Post-training in the wild

# Frontiers labs combine the best of both



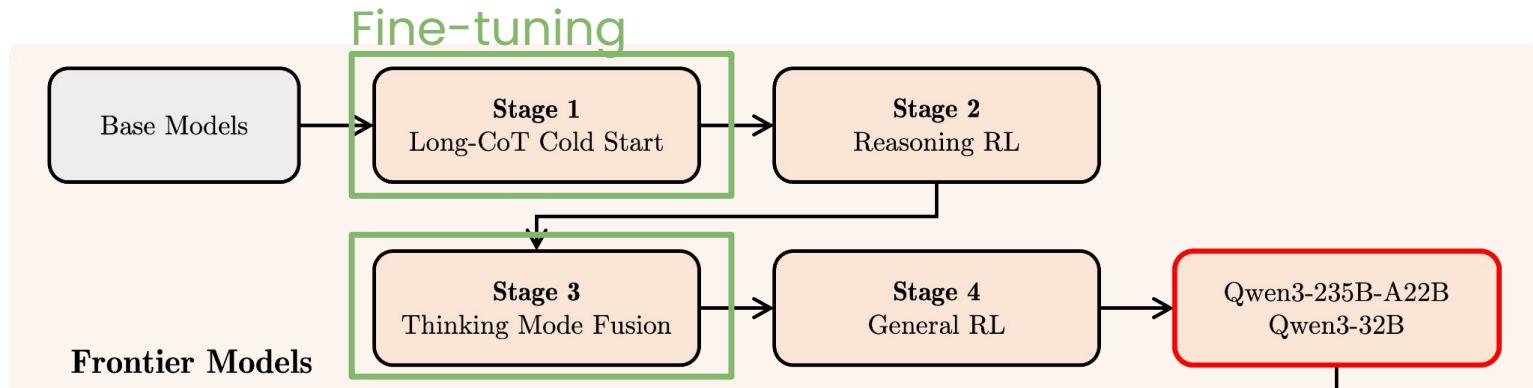
# DeepSeek post-training process

## DeepSeek-R1 Training Pipeline



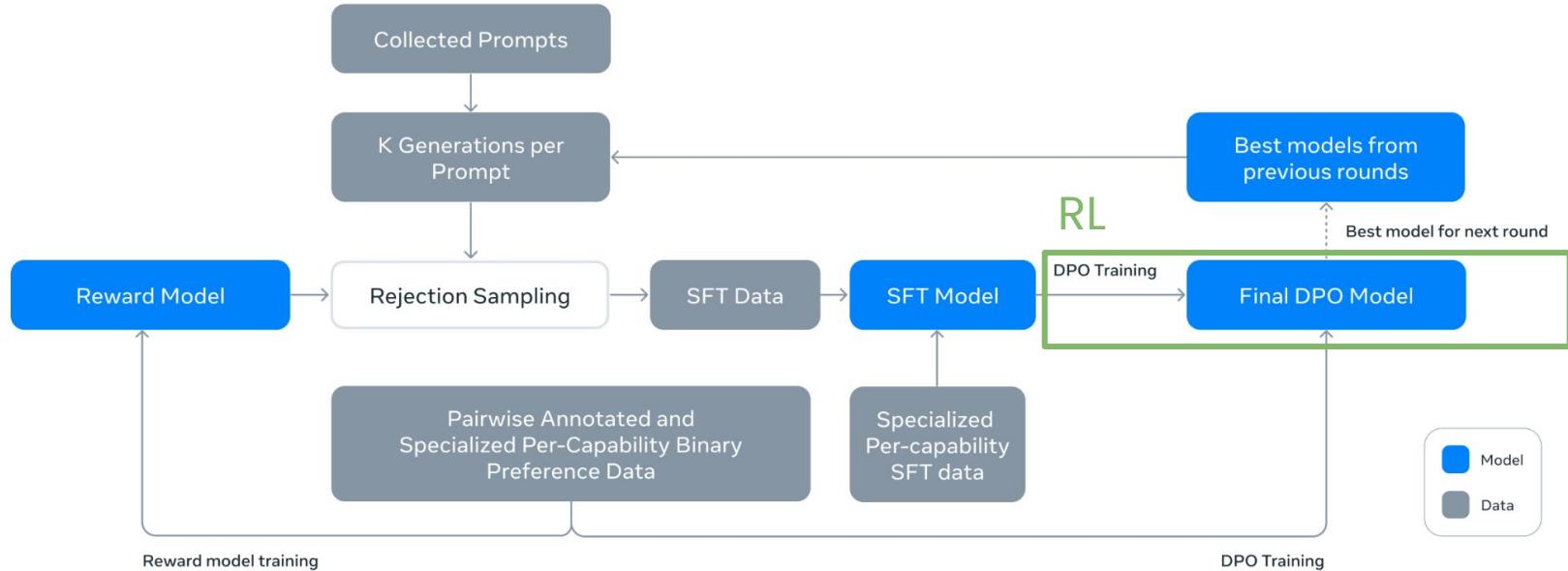
Source: <https://medium.com/@lmpo/deepseek-r1-affordable-efficient-and-state-of-the-art-ai-reasoning-f293b0bd8d65>

# Qwen post-training process



Source: <https://qwenlm.github.io/blog/qwen3/>

# Llama post-training process



*The Llama herd of models. Meta. 2024*

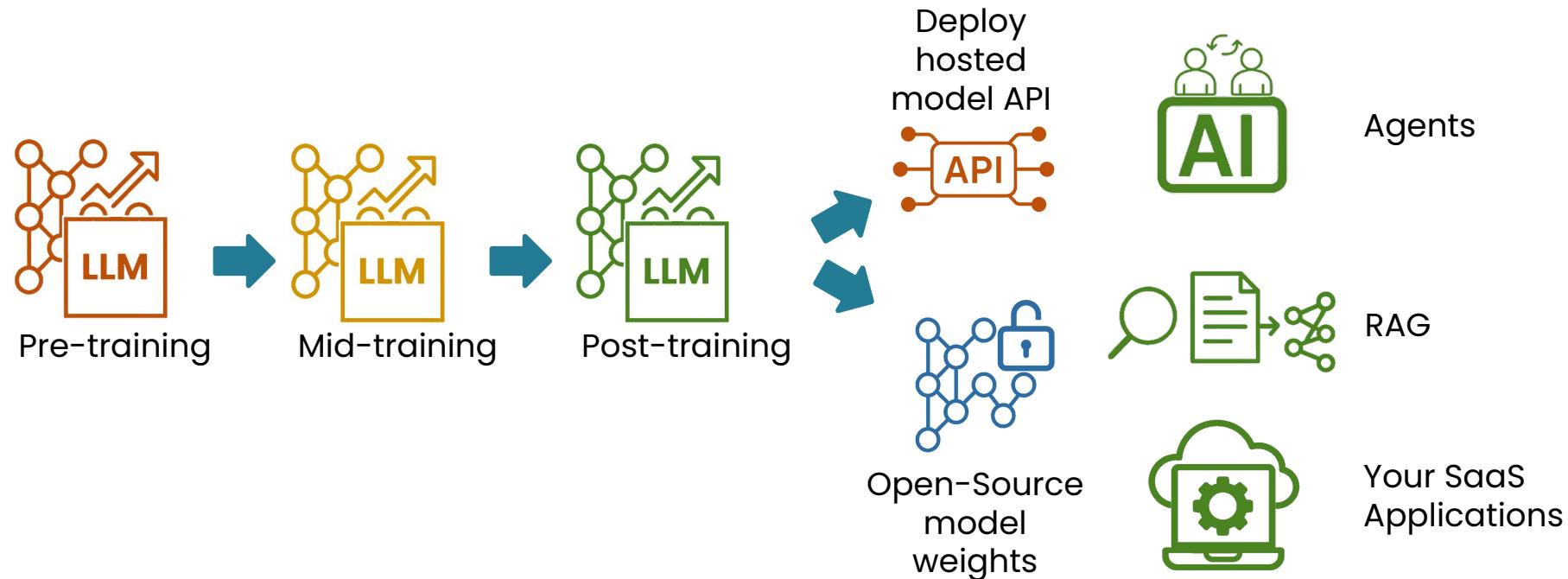
# Where post-training fits in



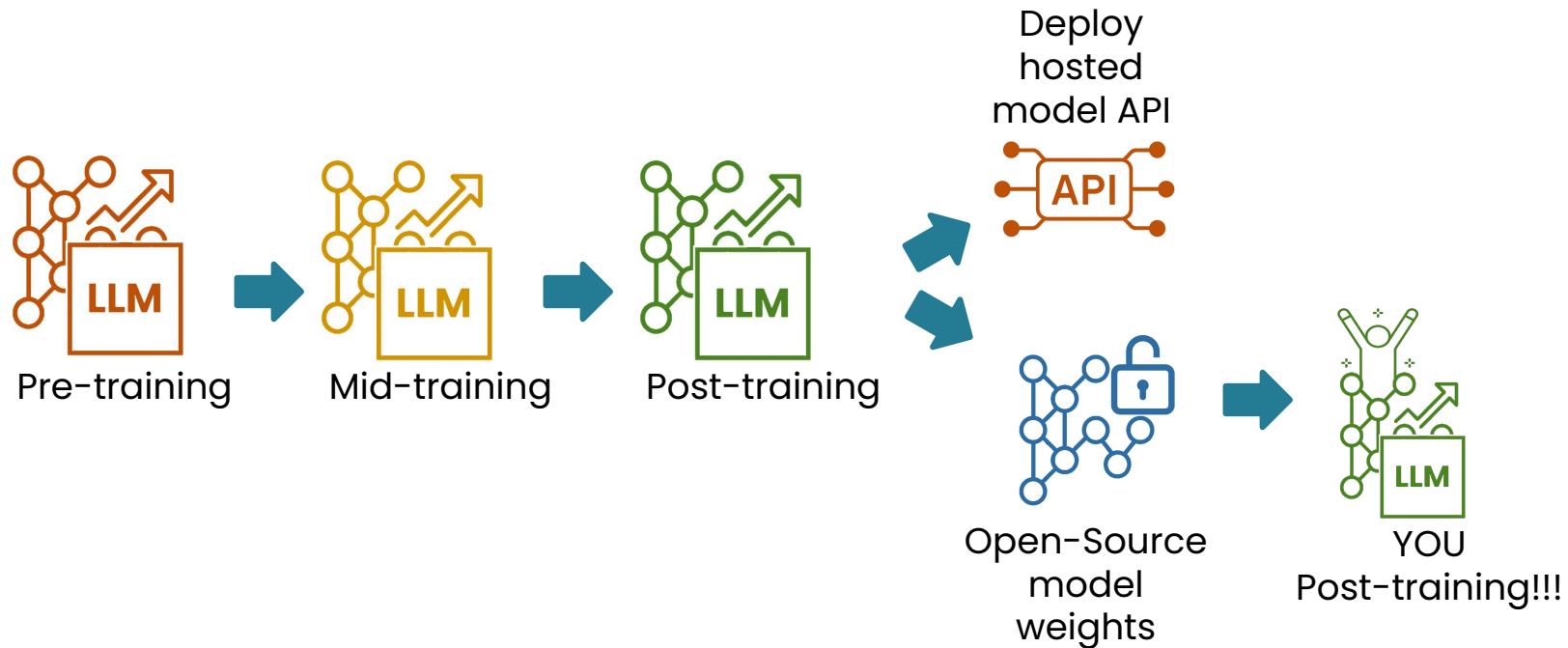
# Where post-training fits in



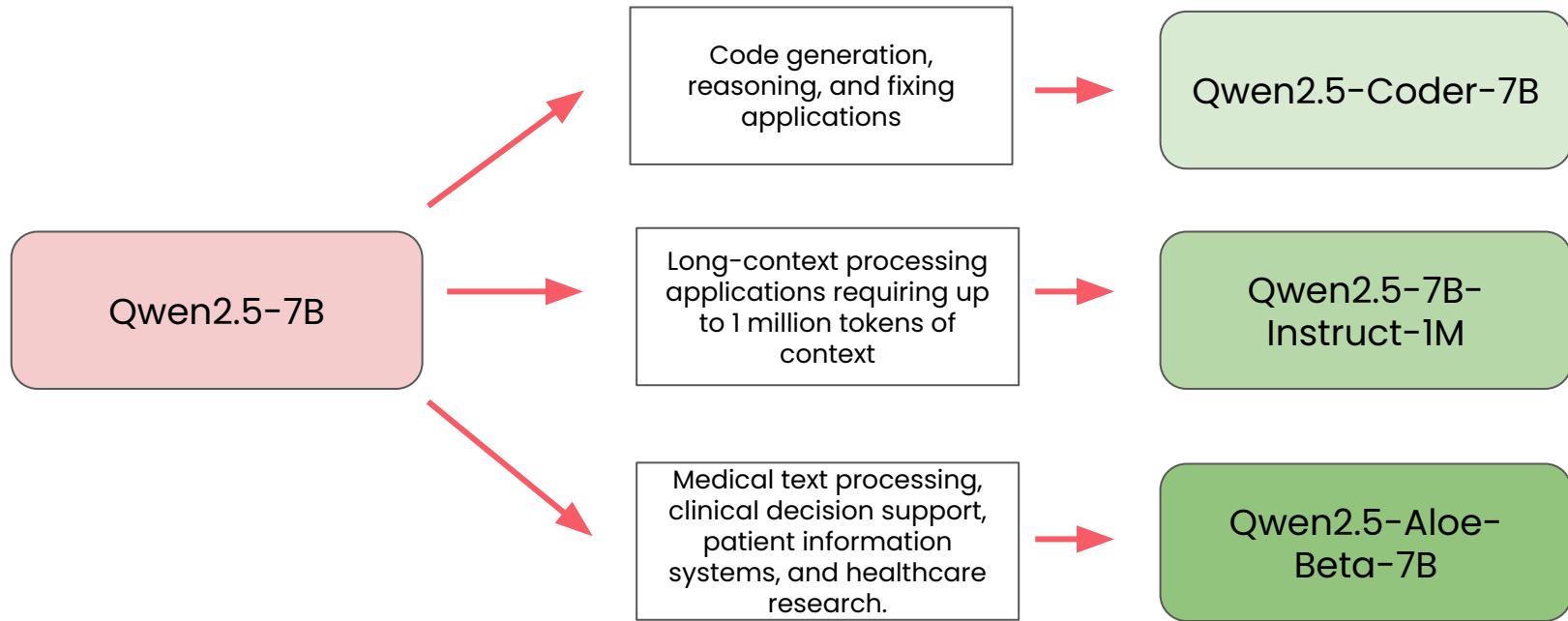
# Where post-training fits in



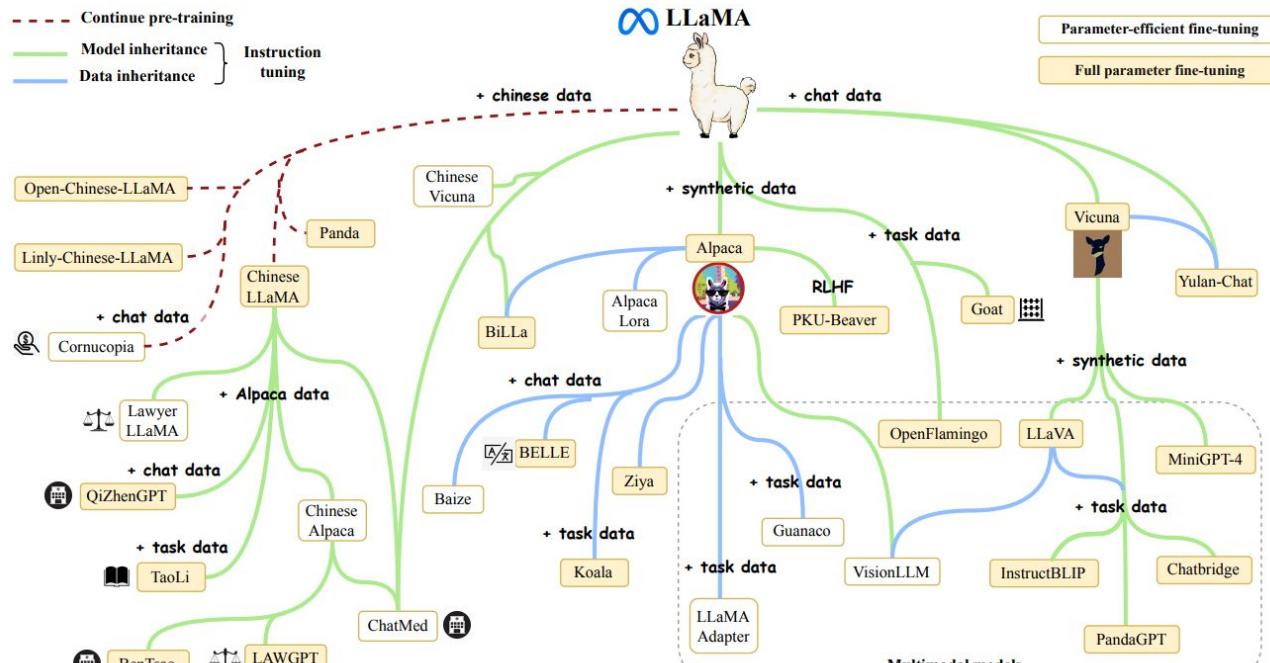
# Where post-training fits in



# Post-training Qwen



# Llama variants



A Survey of Large Language Models. Zhao et al. 2023

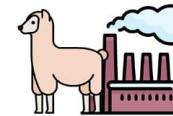
Math Finance Medicine Law Bilingualism Education

# Libraries & services for post-training



**unsloth**

huggingface/trl  
SFT Trainer



**LLaMA-Factory**  
Easy and Efficient LLM Fine-Tuning

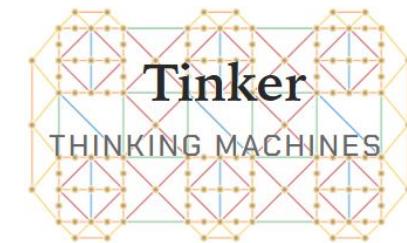
pytorch/  
**torchtune**

**OpenAI**  
Supervised fine-tuning

**mlX**  
Fine-tuning

**together.ai**

**google/tunix**



# In the next modules...

