

ABB - Session 4

Tool use & AI Agents

Shaw Talebi

Today's Session

1. Housekeeping

- 1.1. Announcements
- 1.2. Homework 3

2. AI Agents ↗

- 2.1. What's an AI Agent?
- 2.2. Tool Use
- 2.3. Workflows
- 2.4. Loops & Eval

3. Examples ↗

- 3.1. YouTube Agent with Agents SDK
- 3.2. Upwork Profile Rewriter (in a loop)



Announcements

1) Mid-course Survey

How is the course going so far? *
Select a single answer

1 2 3 4 5 6 7 8 9 10

👎 ----- 😍

What's been working well?

Type your answer

What would you change?

Type your answer

Homework 3

Shoutouts

Persona-based RAG App

Chris Gervais

Duplicate Invoice Detection

Thomas Helms

Economic Profile Generator (HW2)

Ahias Steller Mora

2 Levels of LLM Development

How to get LLMs to do what you want...

Level 1

Adapting models via prompts and tools

Prompt Engineering



RAG



Tool-use



Level 2

Adapting models via additional training

Fine-tuning



Post-training



AI Agents

What are AI Agents?

Something no one can agree on...

“a large language model (LLM), configured with instructions and tools”

“system where a large language model (LLM) can execute more complex tasks through planning and using tools”

“systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks”



Hugging Face

ANTHROPIC

What are AI Agents?

But they do agree on this...

“a **large language model (LLM)**, configured with instructions and tools”

“**system where a large language model (LLM)** can execute more complex tasks through planning and using tools”

“systems where **LLMs** dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks”



Hugging Face

ANTHROPIC

What are AI Agents?

But they do agree on this...

“a **large language model (LLM)**, configured with instructions and **tools**”

“**system where a large language model (LLM)** can execute more complex tasks through planning and using **tools**”

“**systems where LLMs** dynamically direct their own processes and **tool usage**, maintaining control over how they accomplish tasks”



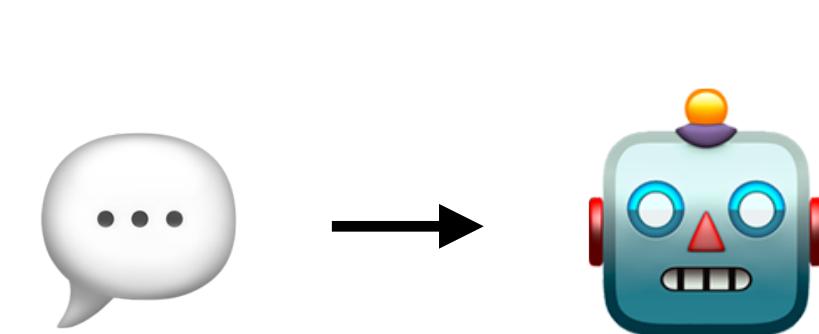
Hugging Face

ANTHROPIC

Agentic Systems

Agency is on a spectrum

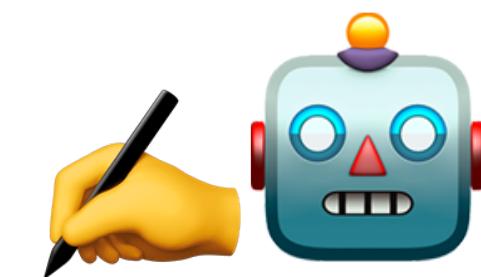
Prompts



Static Prompt



Selects Prompt



Self Prompts



No Agency

Human-level Agency

Agentic Systems

Agency is on a spectrum

Retrieval



Basic Retrieval



Retrieval as a Tool



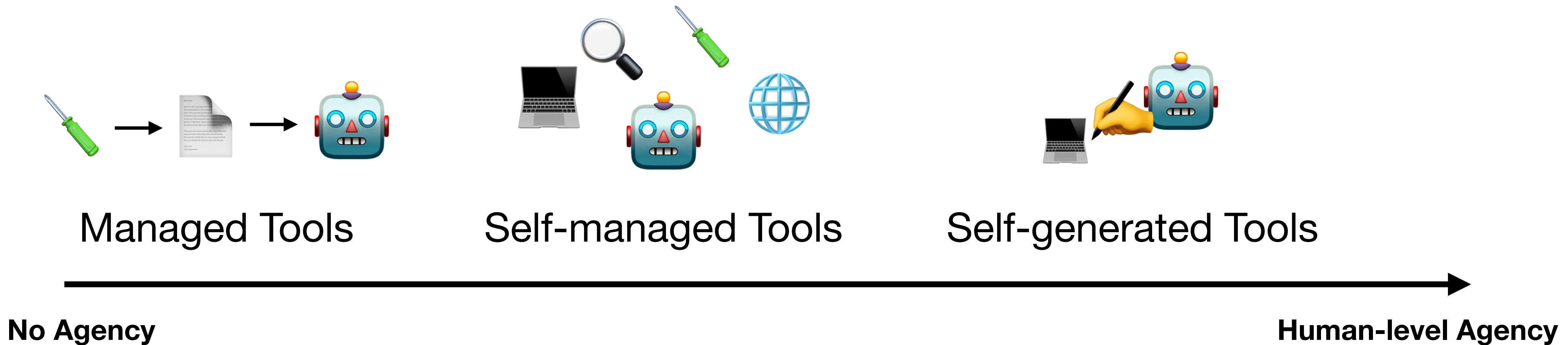
No Agency

Human-level Agency

Agentic Systems

Agency is on a spectrum

Tool Use



Tool Use

Functions that LLMs can call

Simply describe tools

```
<|im_start|>system  
You are a helpful assistant that can use tools when  
needed.  
You have access to the following tool:  
  
- `calculator`: evaluates simple math expressions like  
`5 * 4 + 2`  
  
How to use tools:  
  
If you need to use a tool, respond ONLY in the  
following format:  
  
<tool_call>  
tool_name: calculator  
input: 5 * 4 + 2  
</tool_call>  
  
<|im_end|>
```

***Continuously
monitor outputs
for </tool_call>**

Use fine-tuned model via API

```
{  
  "type": "function",  
  "function": {  
    "name": "calculate",  
    "description": "Performs an arithmetic calculation based on expression.",  
    "parameters": {  
      "type": "object",  
      "properties": {  
        "expression": {  
          "type": "string",  
          "description": "A valid arithmetic expression ex: '3 + 5 * (2 - 8)'."  
        }  
      },  
      "required": [  
        "expression"  
      ],  
      "additionalProperties": false  
    },  
    "strict": true  
  }  
}
```

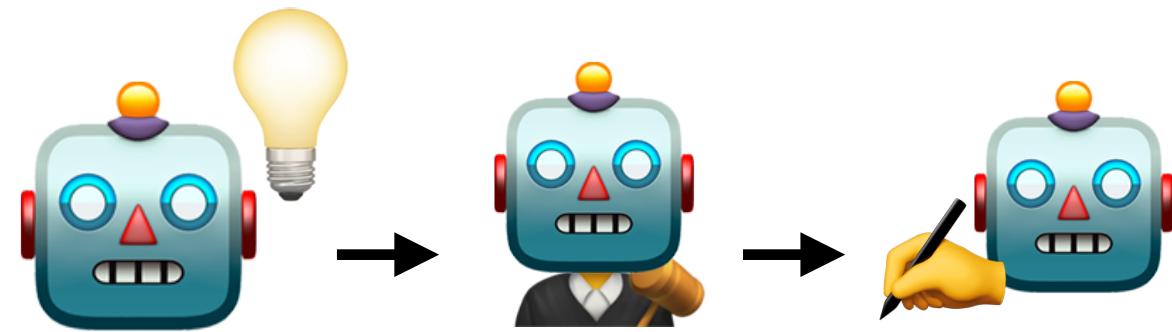


1) API automatically passes tool info model

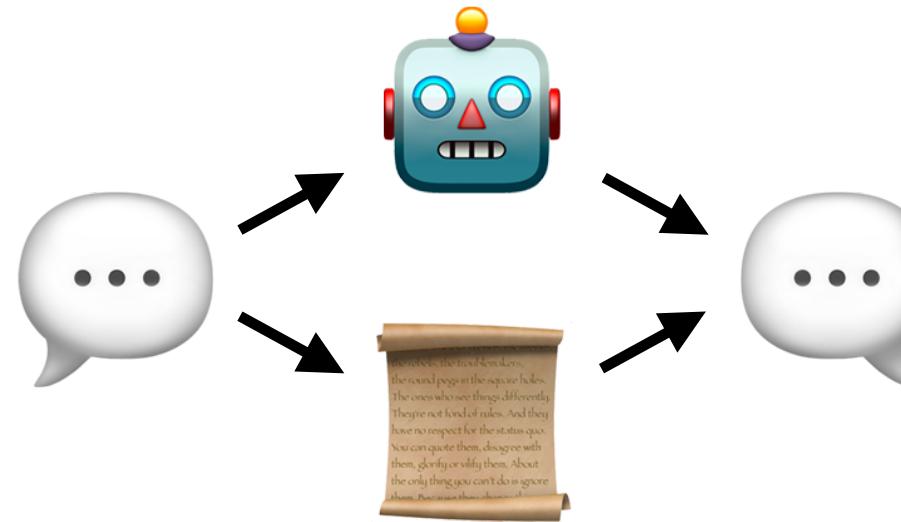
2) API stops inference for tools

LLM Workflows

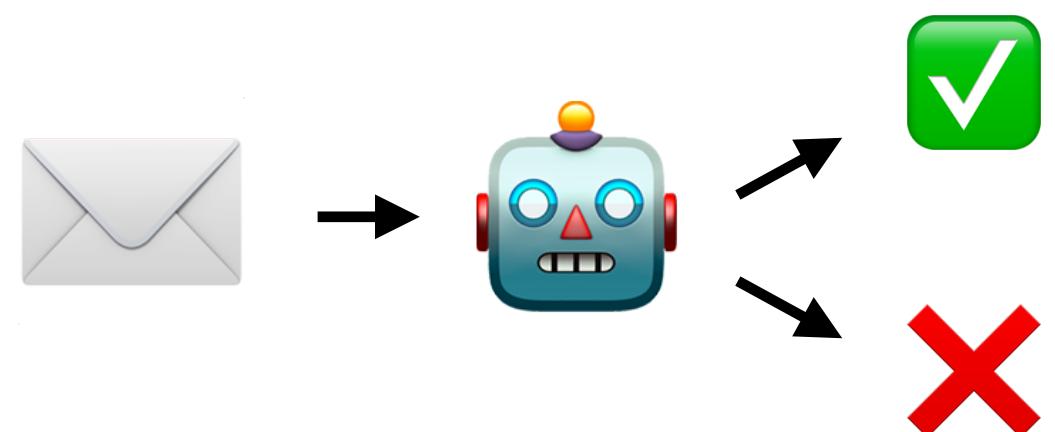
A set of steps that generate a desired result (with at least 1 LLM)



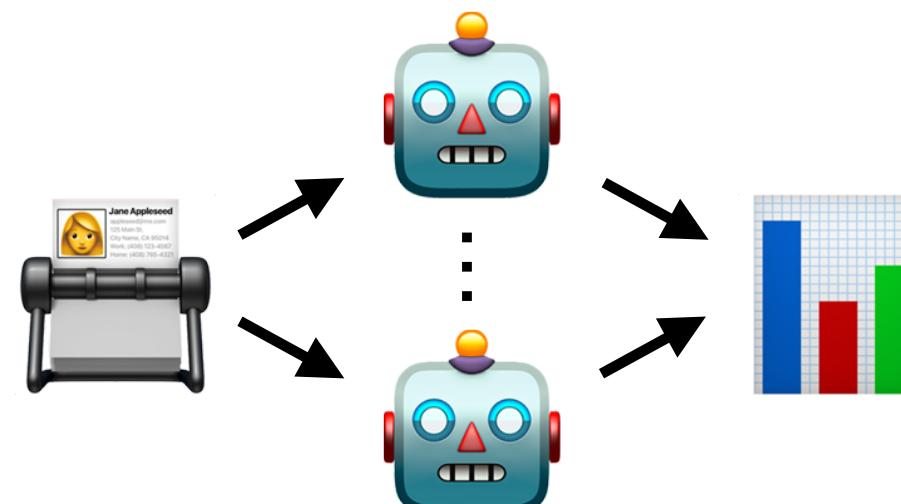
Chaining



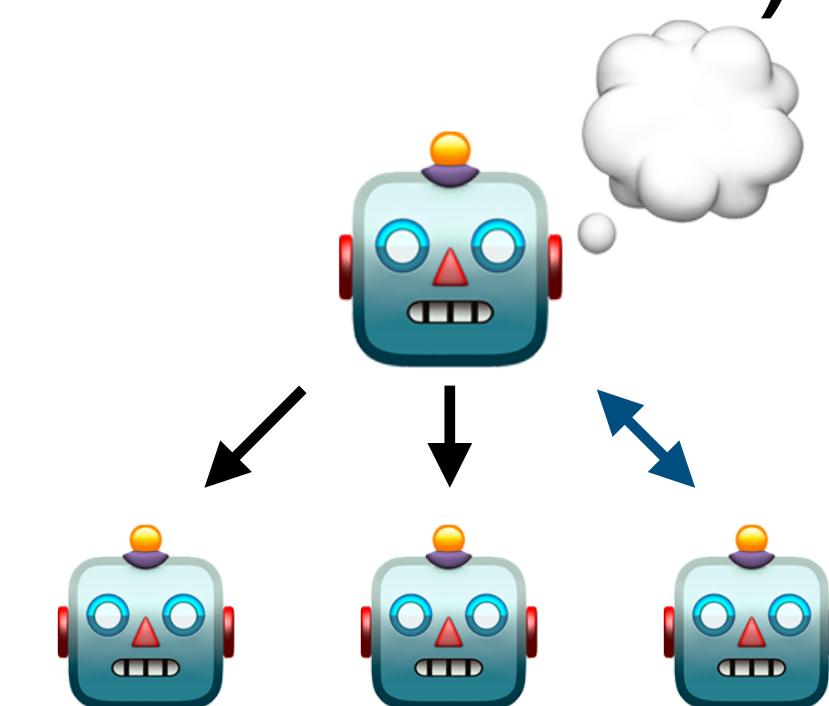
**Parallelization
(Sectioning)**



Routing

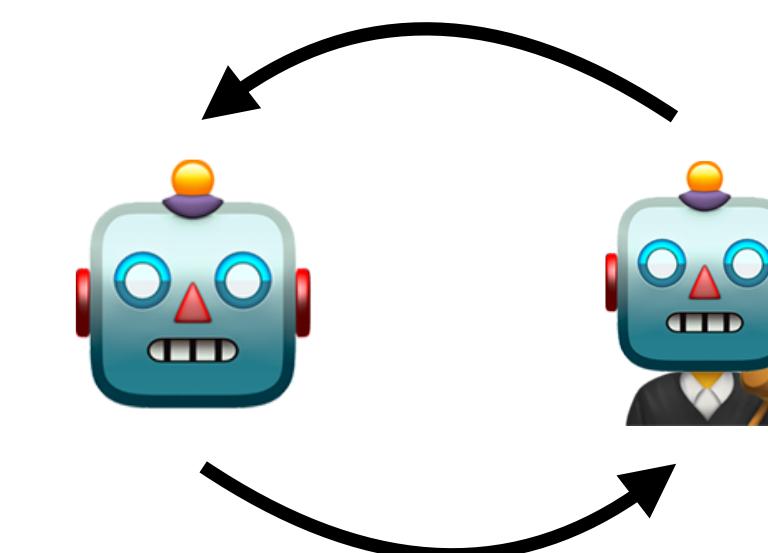


**Parallelization
(Voting)**



(Synthesized Down-flow)

Orchestrator-workers



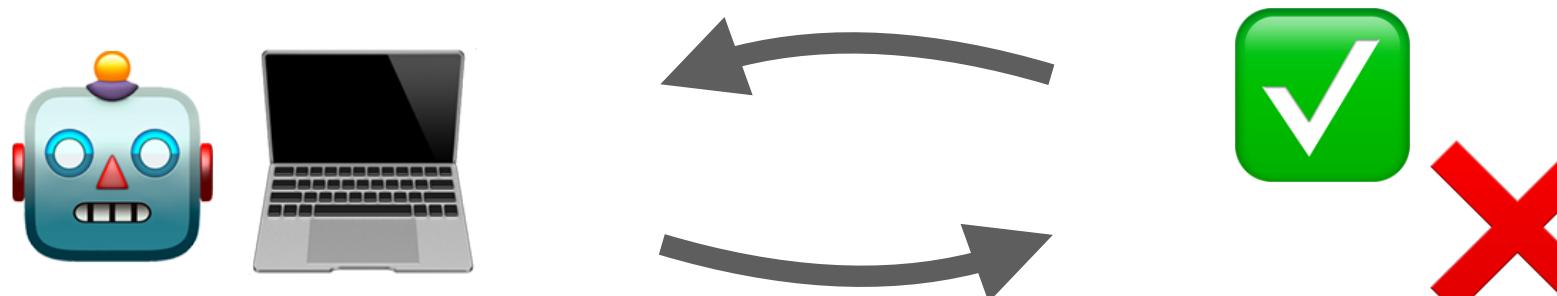
Evaluator-optimizer

LLM-Driven Flow

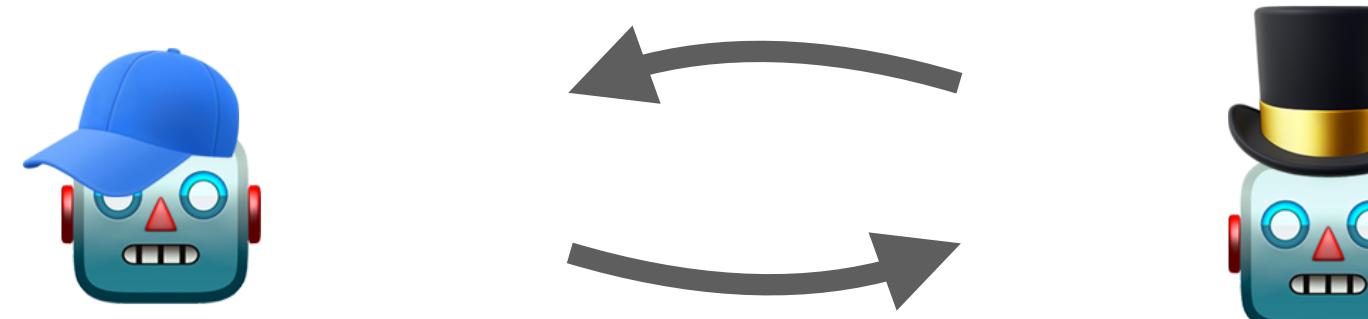
LLM in a Loop

Automatically providing feedback to an LLM to improve outputs

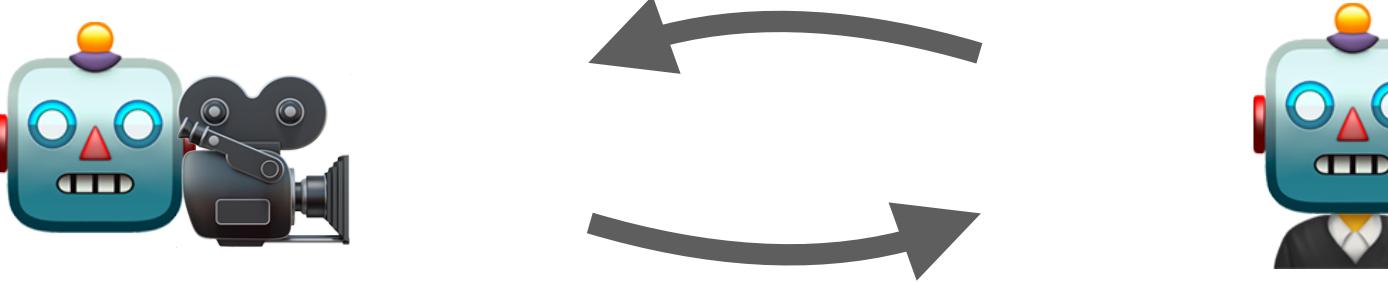
Coding Agent



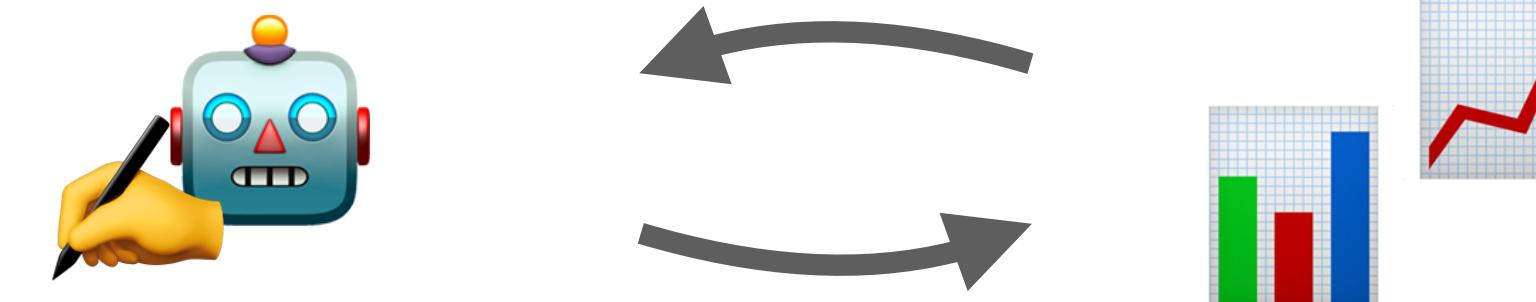
Argument Writer



Video Idea Generator



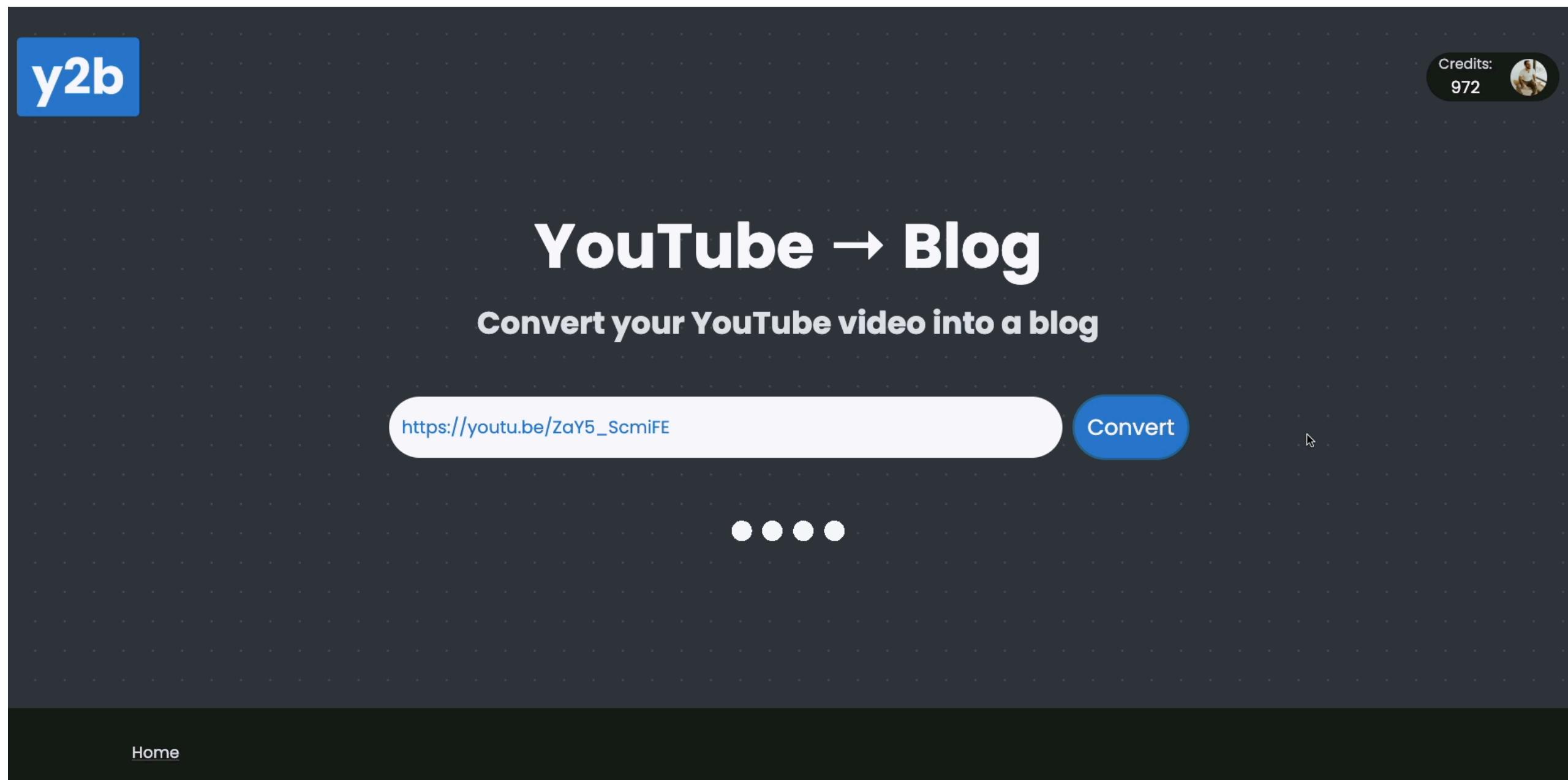
Twitter Agent



Examples

Example 1

YouTube Video Agent (Motivation)

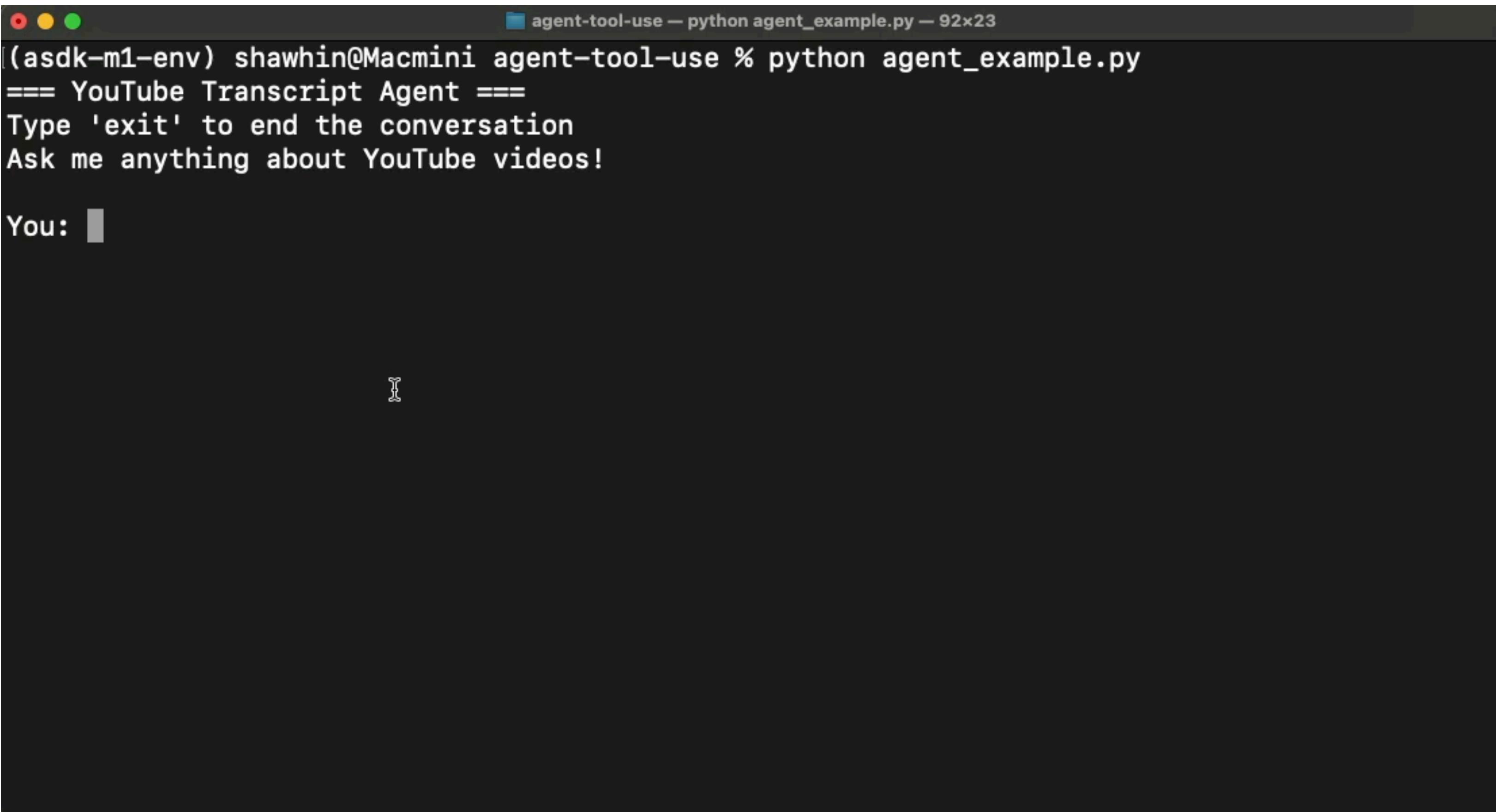


Simple design...

**But limited capability
and flexibility**

Example 1

YouTube Video Agent (Overview)



The screenshot shows a terminal window titled "agent-tool-use — python agent_example.py — 92x23". The window contains the following text:

```
(asdk-m1-env) shawhin@Macmini agent-tool-use % python agent_example.py
== YouTube Transcript Agent ==
Type 'exit' to end the conversation
Ask me anything about YouTube videos!

You: [REDACTED]
```

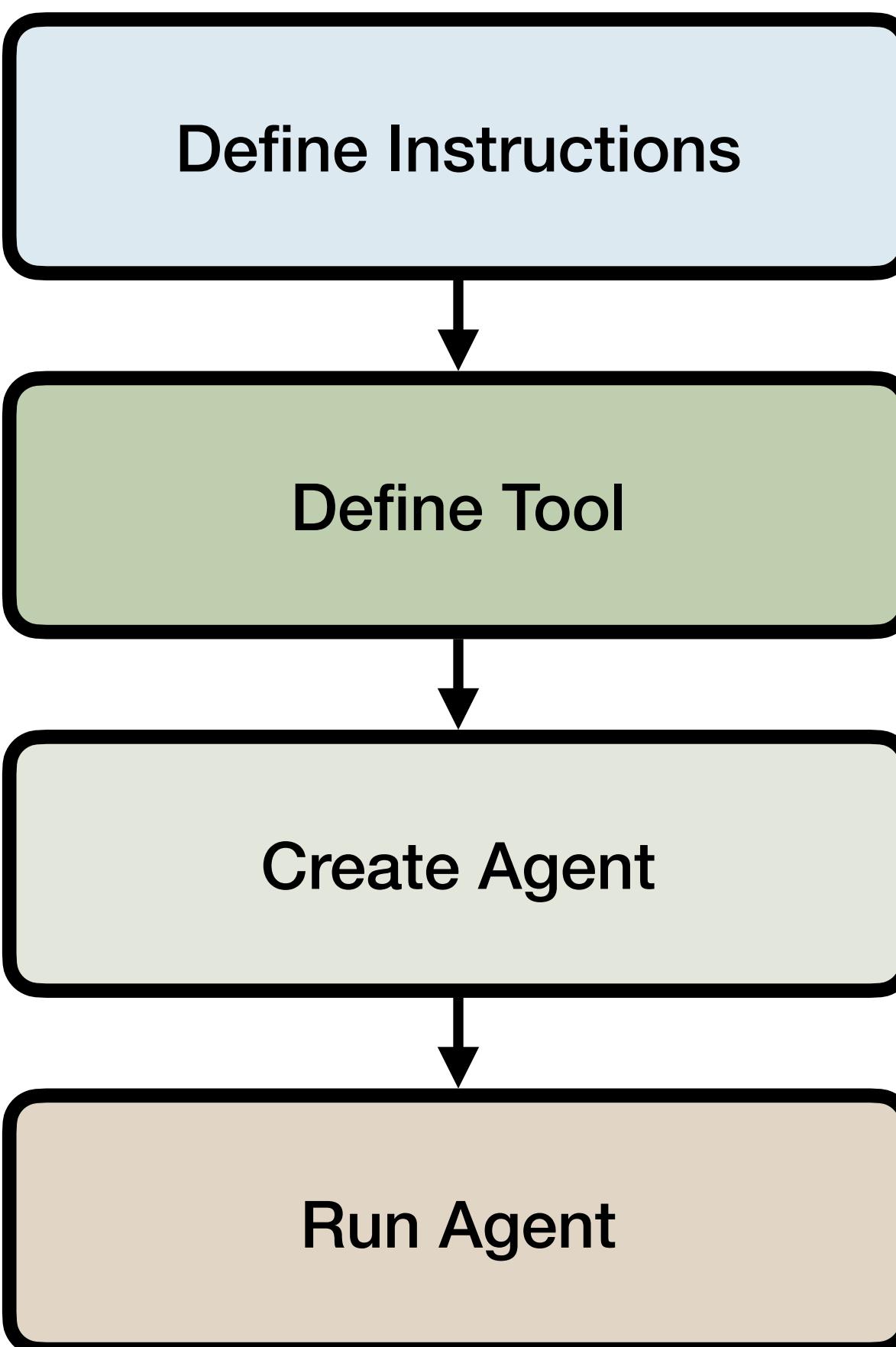
The terminal window has a dark background with light-colored text. The title bar includes the window title and dimensions. The main text area displays the agent's welcome message and instructions. A cursor is visible at the end of the "You:" prompt.

Example 1

YouTube Video Agent (Flowchart)



 OpenAI Agents SDK



Example 1

YouTube Video Agent (Code)



Example 2

Evaluator-Optimizer Profile Rewriter (Motivation)

Data Scientist | PhD, Physics **\$300.00/hr**

Meet Shaw, a data scientist, entrepreneur, researcher, content creator, physicist, drummer, and bread enthusiast driven by an insatiable curiosity to understand the world and create better systems.

With a Ph.D. in Physics from The University of Texas at Dallas, Shaw has over 5 years of experience in data science and project management across diverse fields such as biometric research, credit risk, automotive sales, medicine, human performance, marketing, event planning, and content creation.

As a data scientist, Shaw helps teams solve their problems through data-informed solutions while constantly seeking entrepreneurial opportunities to provide value and tackle challenging problems.

Domain Experience:

- Biometric research
- Credit risk
- Medicine
- Human Performance
- Automotive industry
- Marketing
- Social media analytics
- Content creation



Profile is the main way clients evaluate freelancers

**A bad profile can cost you
2-4X the opportunities**

Example 2

Evaluator-Optimizer Profile Rewriter (Overview)

Data Scientist | PhD, Physics \$300.00/hr

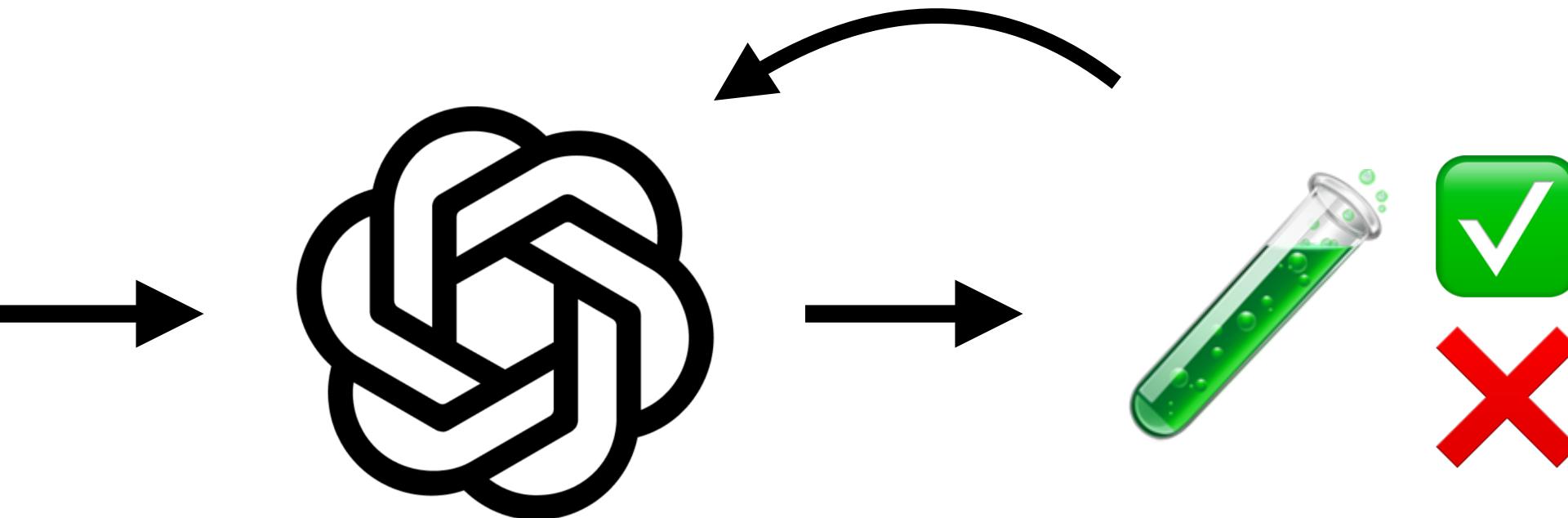
Meet Shaw, a data scientist, entrepreneur, researcher, content creator, physicist, drummer, and bread enthusiast driven by an insatiable curiosity to understand the world and create better systems.

With a Ph.D. in Physics from The University of Texas at Dallas, Shaw has over 5 years of experience in data science and project management across diverse fields such as biometric research, credit risk, automotive sales, medicine, human performance, marketing, event planning, and content creation.

As a data scientist, Shaw helps teams solve their problems through data-informed solutions while constantly seeking entrepreneurial opportunities to provide value and tackle challenging problems.

Domain Experience:

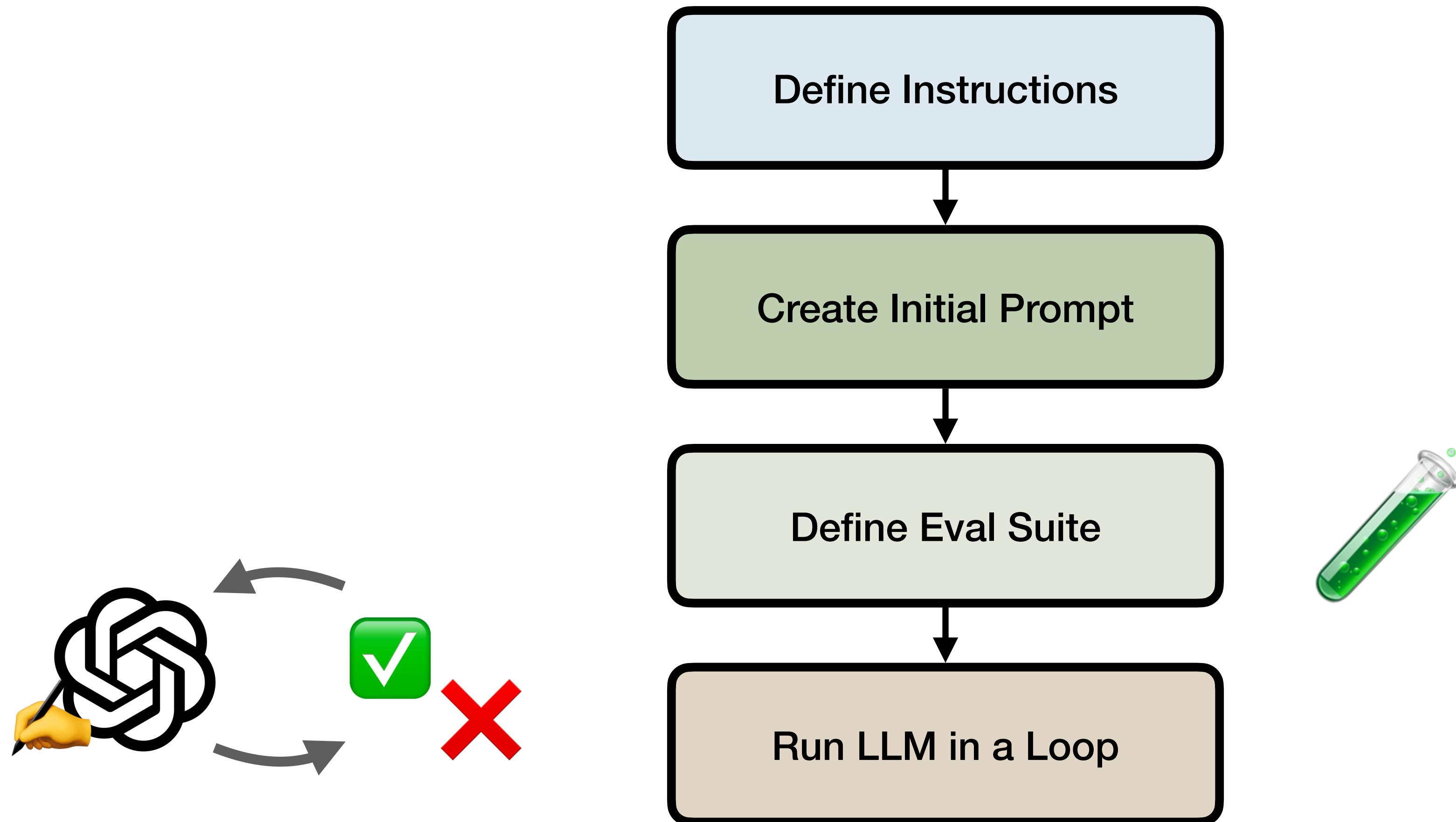
- Biometric research
- Credit risk
- Medicine
- Human Performance
- Automotive industry
- Marketing
- Social media analytics
- Content creation



Test Suite

Example 2

Evaluator-Optimizer Profile Rewriter (Overview)



Homework 4

Project

Build an Agentic System

Pre-work

Session 5: Fine-tuning LLMs

References

- [1] [OpenAI Agents SDK Doc](#)
- [2] [Hugging Face Transformers Doc](#)
- [3] [Building effective agents](#)
- [4] [How to Improve LLMs with Tools](#)
- [5] [LLM Workflows: From Automation to AI Agents](#)
- [6] [arXiv:2402.06627 \[cs.LG\]](#)
- [7] [LLM in a Loop: Automate feedback with evals](#)

Reasoning Models

OpenAI's o1 Model

“Advanced reasoning” model trained on large-scale RL

What is the airspeed velocity of an unladen swallow?

Reasoned for 6 seconds ^

Cheeky question

I'm digging into a Monty Python comedic line about the airspeed velocity of an unladen swallow, where the curious response humorously compares African and European swallows.

Investigating flight dynamics

I'm piecing together the airspeed velocity of an unladen swallow, with a comedic notion of 24 mph. This aligns with Monty Python's playful inquiry, focusing on the African or European species.

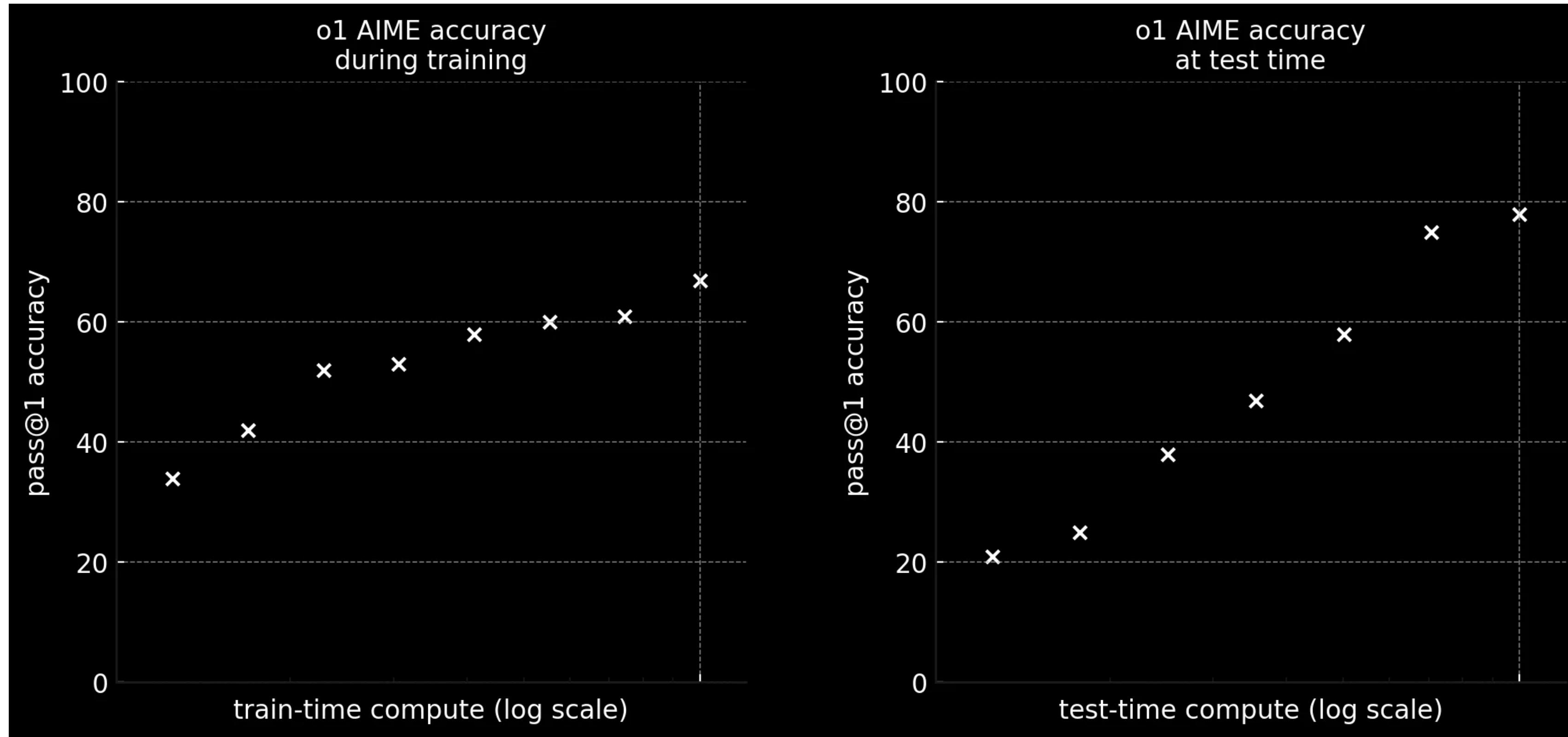
This line, made famous by *Monty Python and the Holy Grail*, is typically answered with, “What do you mean? An African or European swallow?”—leading to King Arthur’s escape from the Bridge of Death when the bridgekeeper cannot reply.

**Model “thinks”
before answering**

Final response

Test-time Compute Scaling

More tokens led to better responses



(Plot reillustrated from [12])

“Thinking” Tokens

Tokens which delimit o1’s chain of thought (CoT)

<think> ... (tokens for thinking through problem) ... </think>

1) Improve UI

Reasoned for 6 seconds ^

Cheeky question

I'm digging into a Monty Python comedic line about the airspeed velocity of an unladen swallow, where the curious response humorously compares African and European swallows.

Investigating flight dynamics

I'm piecing together the airspeed velocity of an unladen swallow, with a comedic notion of 24 mph. This aligns with Monty Python's playful inquiry, focusing on the African or European species.

2) Interpretable Readout

How did they implement this?

DeepSeek Paper

Revealed the mysteries behind o1

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

1) DeepSeek-R1-Zero (RL Only)

Training data for R1

Insights on power of RL

2) DeepSeek-R1(SFT + RL)

Better performance

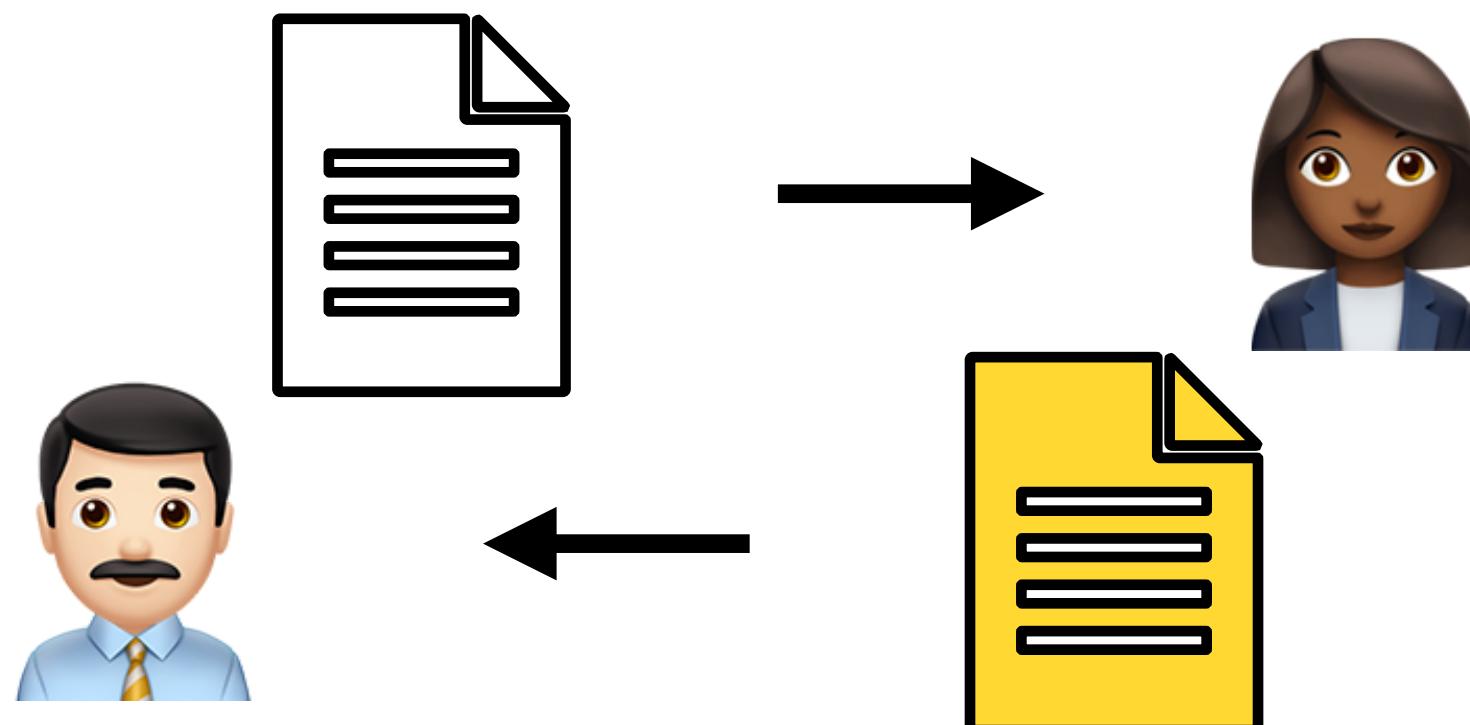
More public attention

DeepSeek-R1-Zero (RL Only)

RL = Reinforcement Learning

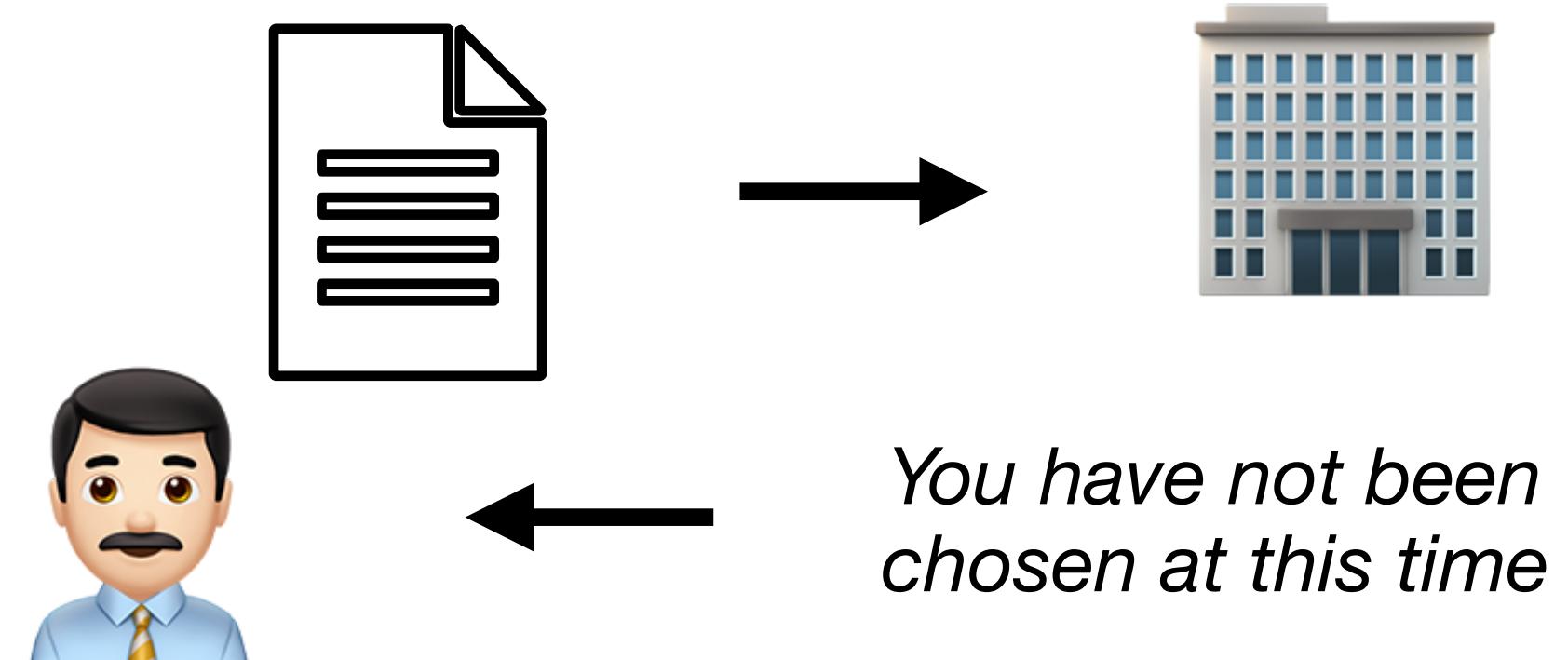
Supervised Learning

Model learns by example



Reinforcement Learning

Model learns by trial and error

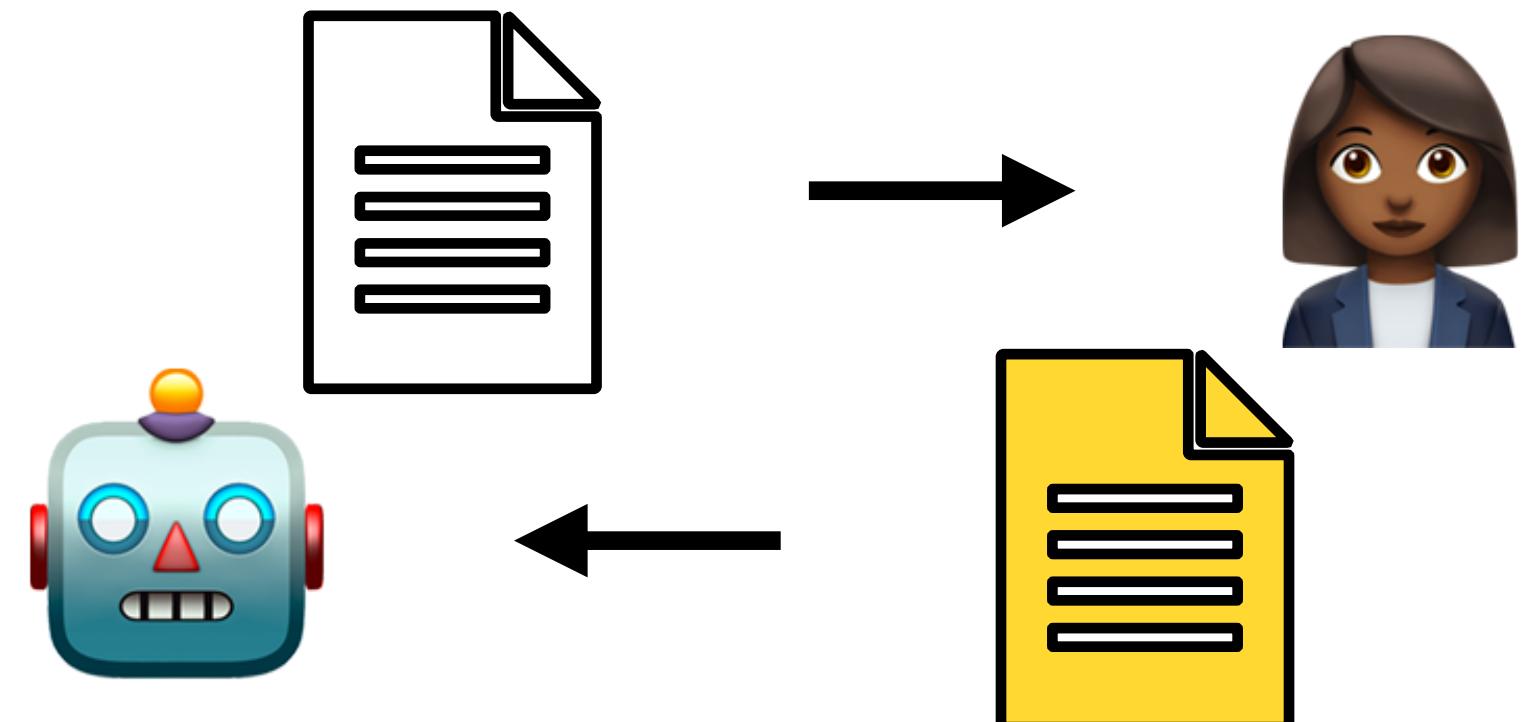


DeepSeek-R1-Zero (RL Only)

RL = Reinforcement Learning

Supervised Learning

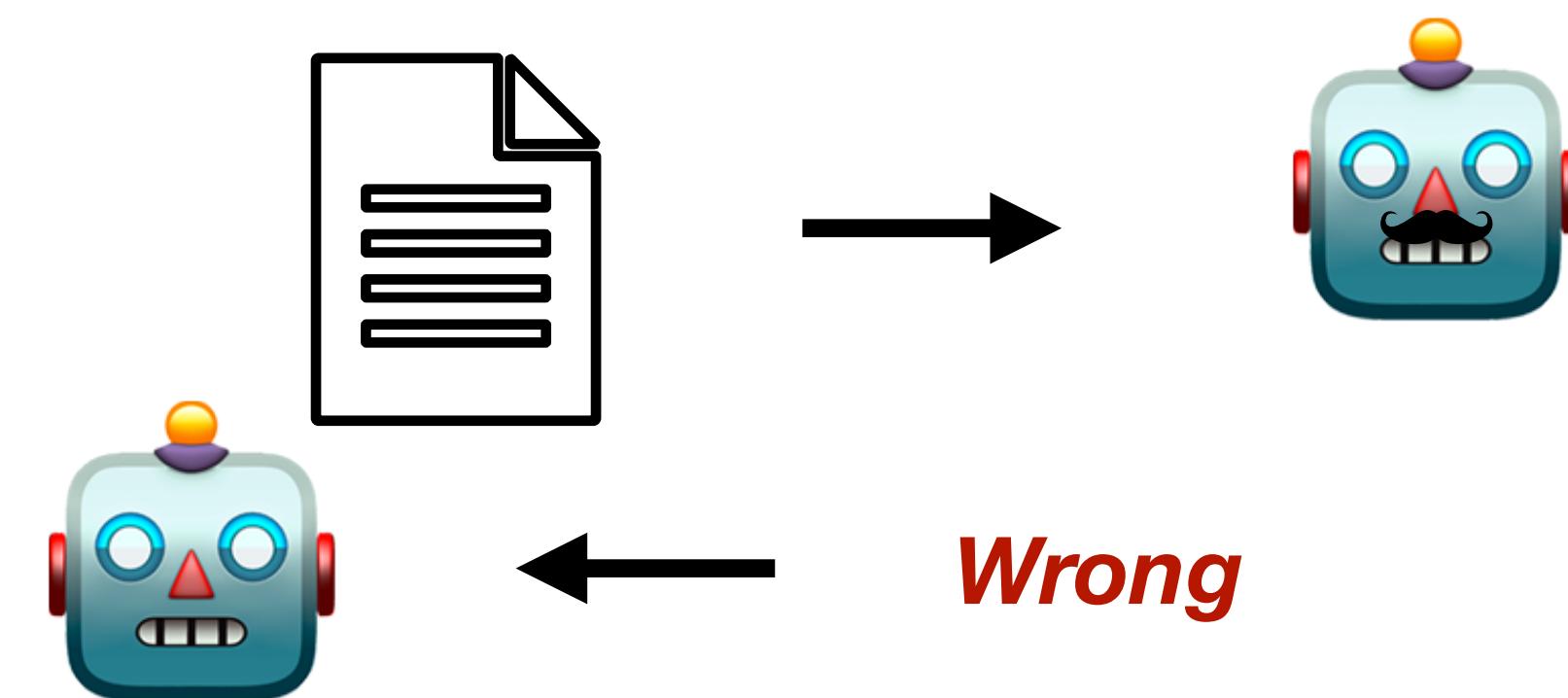
Model learns by example



"Make it more like this"

Reinforcement Learning

Model learns by trial and error



Rule-based Reward

The incentive we give to the model to do what we want

1) Accuracy

A 5×100 table is divided into 500 unit square cells, where n of them are colored black and the rest are colored white. Two unit square cells are called adjacent if they share a common side. Each of the unit square cells has at most two adjacent black unit square cells. Find the largest possible value of n

302

Right/wrong check

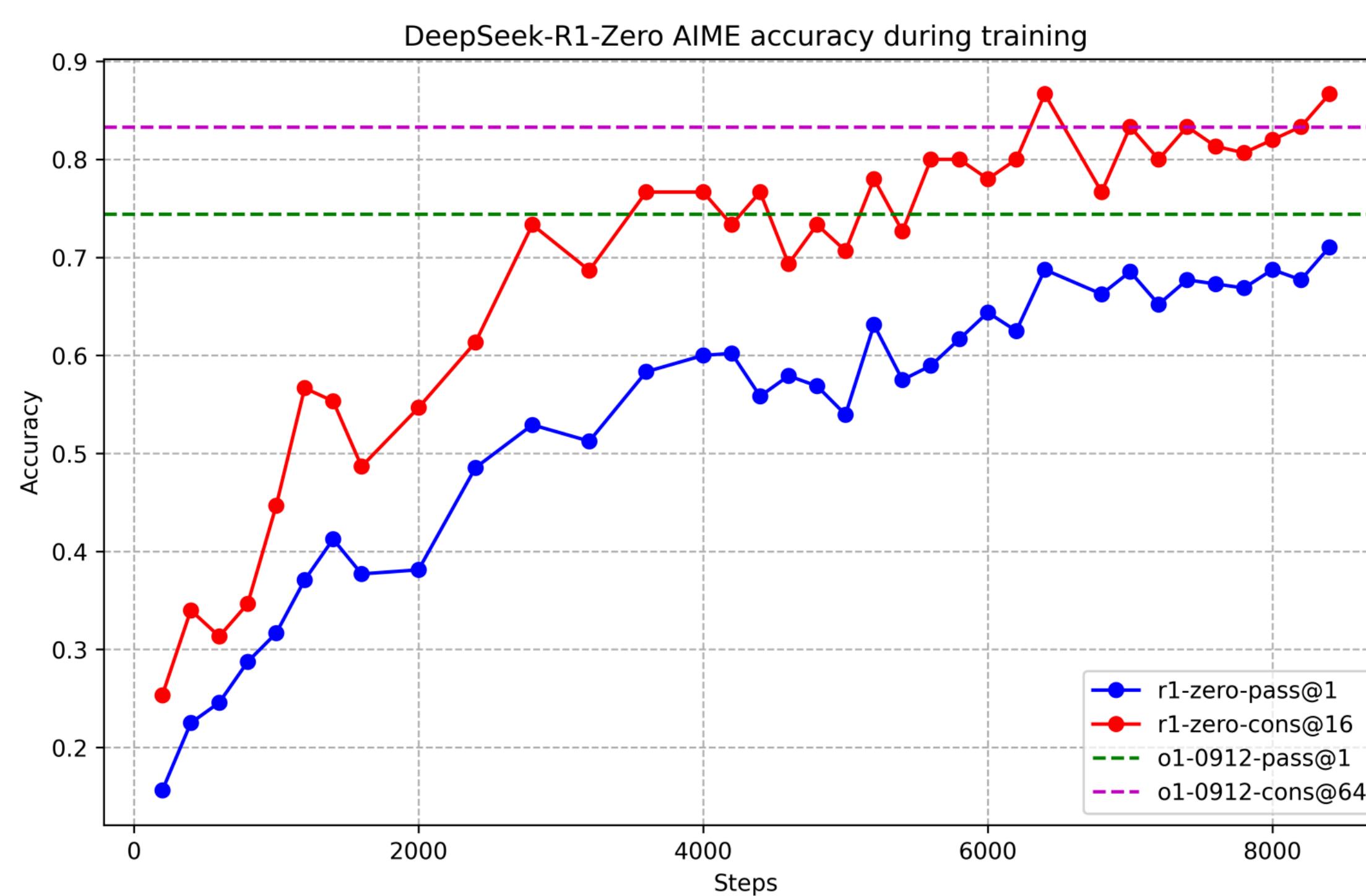
2) Formatting

<think> Okay, so I need to figure out the maximum number of black cells (n) in a 5×100 grid where each cell can have at most two adjacent black cells... </think>

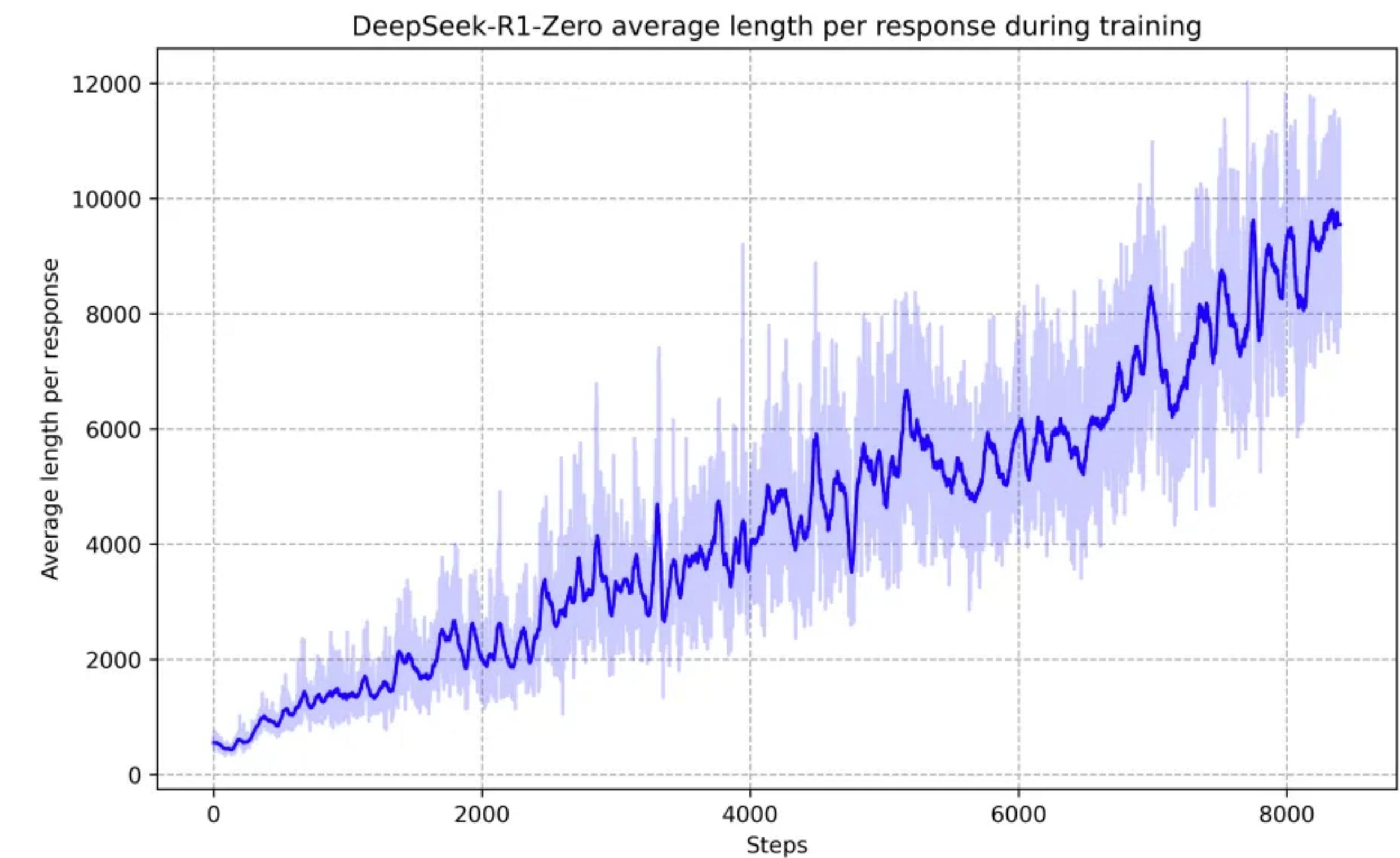
Check if reasoning is within think tags

Results

Performance and test-time compute



(Fig 2 from [13])



(Fig 3 from [13])

Results

Emergent abilities: Long CoT & Reflection

User prompt: If it takes me 1.5 hours to write a codebase and my intern 8 hours. How long will it take us both to write 3 codebases?

<think> To solve for how long it will take both you and your intern to write three codebases together, we need to first find out how many codebases each of you can write per hour individually.

(long CoT) ...

Long CoT

So, it will take approximately 3.7895 hours for you and your intern to write three codebases together.

To make sure everything makes sense, let us round the answer to a reasonable number of decimal places (e.g., round to two decimal places):

$t \approx 3.79$ hours

However, if you need a more "human-friendly" answer, you could convert the decimal part into minutes. Since 0.7895 of an hour is:

$0.7895 \times 60 \text{ minutes} \approx 47.37 \text{ minutes}$

So, 3.7895 hours is approximately 3 hours and 47.37 minutes (or roughly 3 hours and 47 minutes if rounding to the nearest minute).

Verification (optional step for verification)
To verify, let us check how many codebases you can individually write in $t = 72/19$ hours:

$$\begin{aligned} &\text{Your work in } 72/19 \text{ hours} \\ &= (2/3 \text{ codebases per hour}) \times (72/19 \text{ hours}) \\ &= (2 \times 72) / (3 \times 19) \\ &= 144 / 57 \\ &= 48/19 \text{ codebases.} \end{aligned}$$

(verification process) ...

However, a more "exact" answer should be kept in fractional form unless specified otherwise. Thus, the answer is: $72/19$ hours which is approximately \$3.7895\$ hours or about \$3.79\$ hours when rounded to two decimal places).
</think>

Verification

72/19 hours (or approximately 3.79 hours).

Evals

Vibe Checks

Evaluating performance by skimming single examples

Write me a YouTube title for a video on LLM evals.

"Mastering LLM Evaluations: Key Metrics and Best Practices"

Write me a YouTube title for a video on LLM evals. Don't start with vague words like "unlocking" or "mastering"

"How to Effectively Evaluate Large Language Models: Key Metrics and Methods"

Vibe Checks

Evaluating performance by skimming single examples

Write me a YouTube title for a video on LLM evals. Don't start with vague words like "unlocking" or "mastering". Make sure it's no longer than 65 characters. Shorter is better.

"How to Evaluate LLMs for Real-World Applications"



Helps you move fast



Great way to go from 0 to 1



Comparing multiple designs is a pain



Leaves performance on the table

Evals

A number that correlates with the outcome you care about

	<u>Length</u>	<u>CTR</u>	<u>Expert Grade</u>
Write me a YouTube title for a video on LLM evals. "Mastering LLM Evaluations: Key Metrics and Best Practices"	59	2%	Fail
Write me a YouTube title for a video on LLM evals. Don't start with vague words like "unlocking" or "mastering" "How to Effectively Evaluate Large Language Models: Key Metrics and Methods"	76	3%	Fail
Write me a YouTube title for a video on LLM evals. Don't start with vague words like "unlocking" or "mastering". Make sure it's no longer than 65 characters. Shorter is better. "How to Evaluate LLMs for Real-World Applications"	50	4.2%	Pass

3 Types of Evals

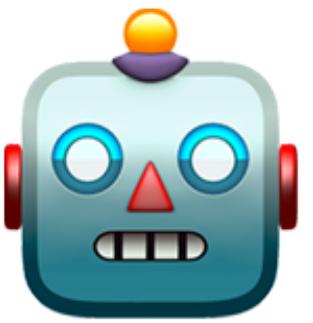
Type 1: Code-based



Type 2: Human-based



Type 3: LLM-based



Type 1: Code-based

Rule-based assessment of model's output

$$2 + 2 = 4 \quad \checkmark$$

$$812 \times 431 = 346,772 \quad \times$$

Who was the 35th president of the US?

- A) JFK
- B) Lyndon Johnson
- C) Donald Trump
- D) FDR

A 

Correct Answer

Formatting Checks

NLP Metrics

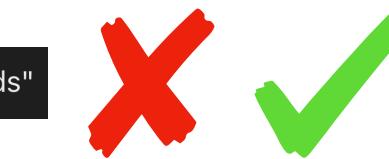
 < 65 chars

 Doesn't start with X-ing

"Mastering LLM Evaluations: Key Metrics and Best Practices"



"How to Effectively Evaluate Large Language Models: Key Metrics and Methods"



"How to Evaluate LLMs for Real-World Applications"

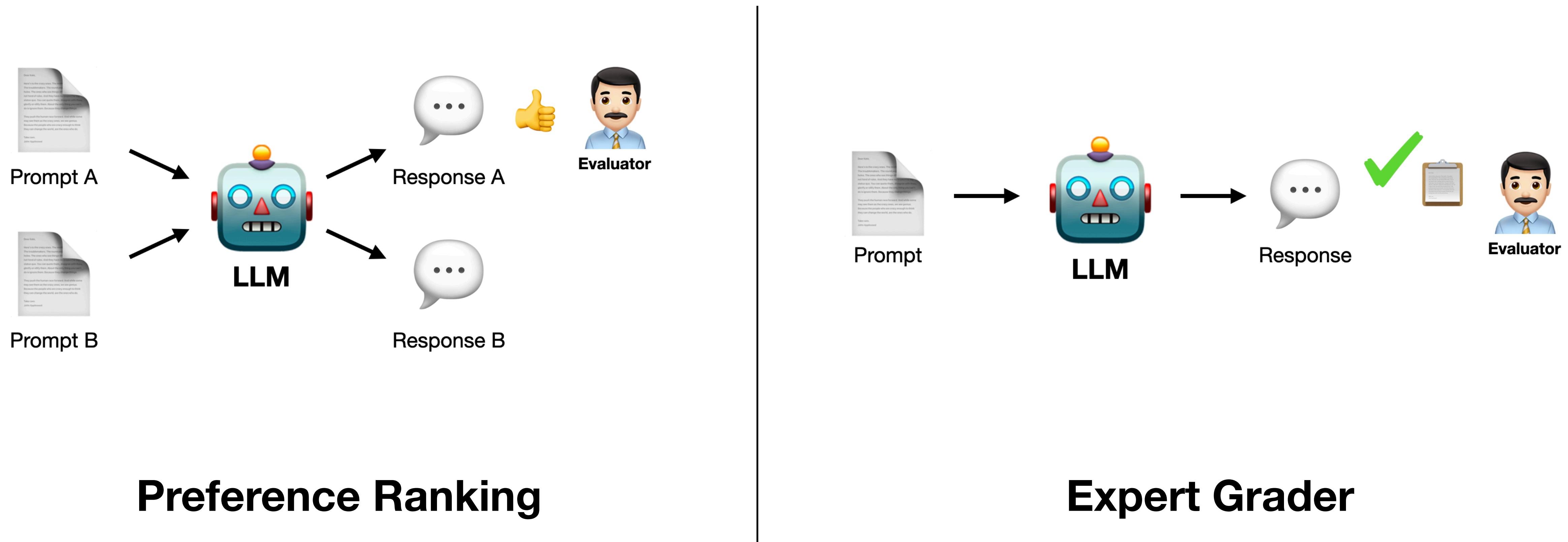


 BLEU (translation)

 ROUGE (summarization)

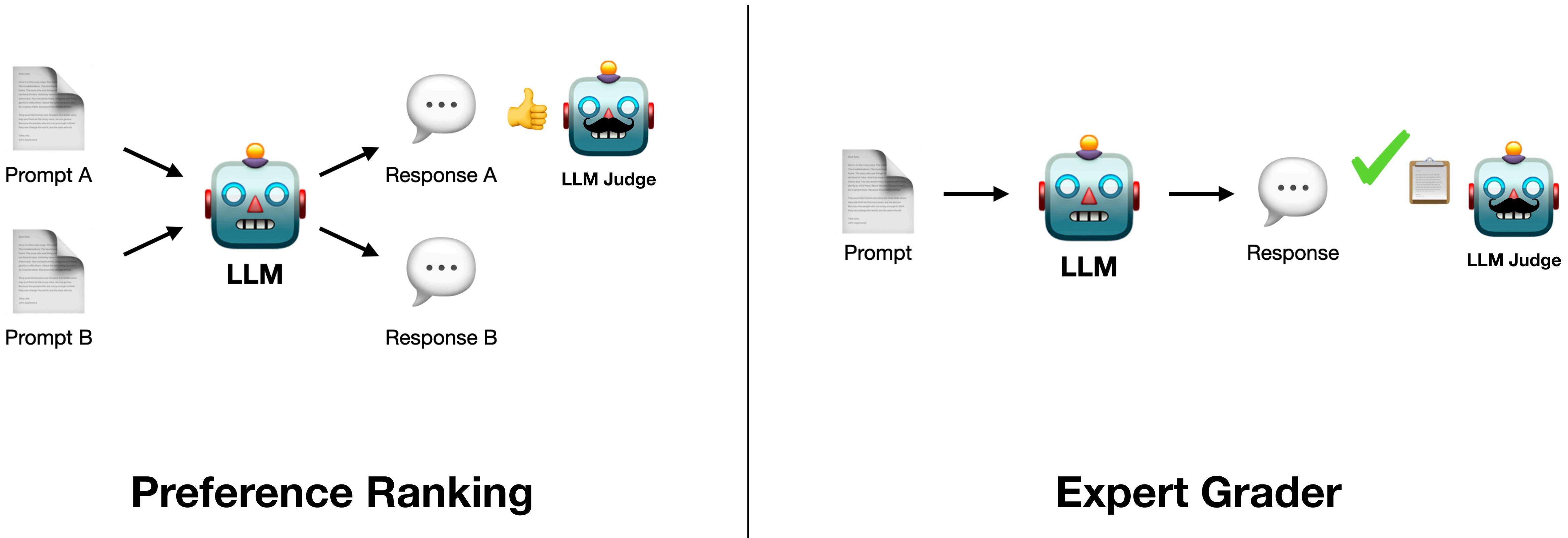
Type 2: Human-based

Using human judgement (or behavior) assess output quality



Type 3: LLM-based

Using an LLM to evaluate output quality



Typically via Prompt Engineering
... but can also be done through training