

Building AI Assistants in an Evolving Landscape

Ashwin Ramesh
ML Engineer

About Me

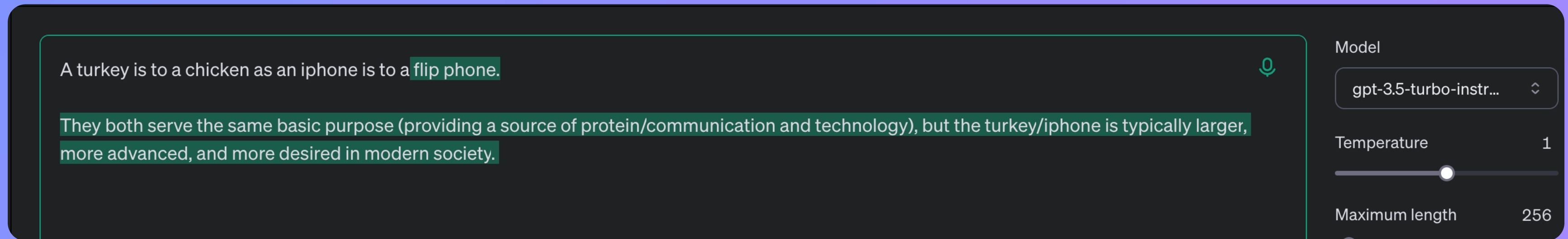
- MS CS@Illinois '20
- DL Inference @ NVIDIA
- Reasoning & Planning for Copilots @ Continual AI



Outline

- How we are using LLMs today
- Advanced LLM Capabilities
- The next-generation of AI Assistants

LLMs are Scaled Next Token Predictors



A turkey is to a chicken as an iphone is to a flip phone.

They both serve the same basic purpose (providing a source of protein/communication and technology), but the turkey/iphone is typically larger, more advanced, and more desired in modern society.

Model: gpt-3.5-turbo-instr...
Temperature: 1
Maximum length: 256



And then there was chat...

LLMs Are Knowledge Bases



You

What's the linux command to kill a process at a port

...

Show drafts

✦ There are two main commands you can use to kill a process based on the port it's using on Linux:

1. **fuser**: This command identifies and manipulates processes using specific files and file systems.
2. **lsof**: This command lists information about open files.

Here's how to use each command:

Using fuser:

- **Identify the process:**
 - Run `fuser -n tcp <port_number>` (replace `<port_number>` with the actual port).
 - This will display the PID (Process ID) of the process using that port (if any).
- **Kill the process:**
 - Run `sudo kill <PID>` (replace `<PID>` with the PID you obtained from the

You

Explain the node.js. programming model in the style of a 90's rap song

ChatGPT

Yo, yo, check it, let's dive into Node.js, keeping it tight,
With a flow from the 90's, we're gonna light up the night.

- **Verse 1:**

Yo, Node.js in the scene, it's the runtime dream,
Built on Chrome's V8, it's fast, makes your code gleam.
Event-driven, non-blocking I/O is the theme,
Makes concurrent execution slick, like it's part of the team.
- **Chorus:**

Node.js, oh Node.js, with your model so clean,
Async operations, on the JavaScript machine.
No waiting, just staving, events keep it keen,
Node.js, oh Node.js, making performance supreme.
- **Verse 2:**

Now, let's talk about the core, the Event Loop's lore,
Single-threaded magic, but it handles much more.
It queues up callbacks, doesn't make your CPU sore,
Handles thousands of connections, without needing to roar.
- **Bridge:**

LLMs Are Reasoning Engines

You
Here's the output of git reflog. I want to undo the latest git reset

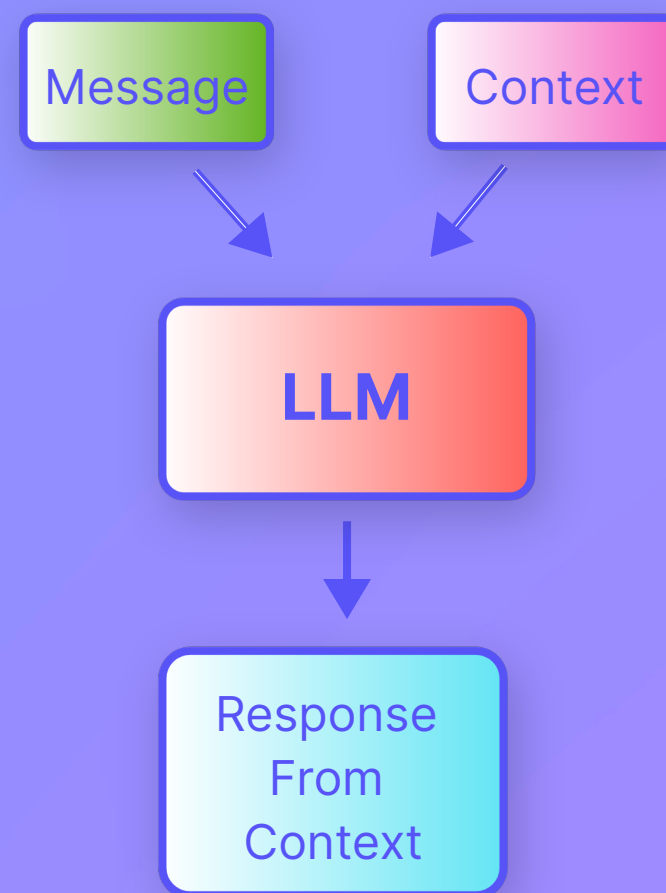
```
c54d9268 (HEAD -> ashwinramesh/dev-5509) HEAD@{0}: reset: moving to HEAD-2
3f411213 HEAD@{1}: rebase -i (abort): updating HEAD
c54d9268 (HEAD -> ashwinramesh/dev-5509) HEAD@{2}: rebase -i (start): checkout HEAD-10
3f411213 HEAD@{3}: commit: Removed old service files
8584fdd9 HEAD@{4}: pull origin ashwinramesh/dev-5509: Merge made by the 'recursive'
strategy.
c54d9268 (HEAD -> ashwinramesh/dev-5509) HEAD@{5}: checkout: moving from
ashwinramesh/dev-5599 to ashwinramesh/dev-5509
4a7a0109 (ashwinramesh/dev-5599) HEAD@{6}: commit: refactoring workflow
c80d0ffa HEAD@{7}: rebase finished: returning to refs/heads/ashwinramesh/dev-5599
c80d0ffa HEAD@{8}: pull --rebase origin main: Engine type
954dc778 HEAD@{9}: pull --rebase origin main: Refactoring workflows
10afd7b2 HEAD@{10}: pull --rebase origin main: Refactoring workflows
ca865bf8 (origin/main, origin/HEAD) HEAD@{11}: pull --rebase origin main: checkout
ca865bf8e04806f846fb2e0c5944cb5e1fa75fdf
d6b43402 HEAD@{12}: commit: Engine type
657c989b HEAD@{13}: rebase finished: returning to refs/heads/ashwinramesh/dev-5599
657c989b HEAD@{14}: pull --rebase origin main: Refactoring workflows
c324170b HEAD@{15}: pull --rebase origin main: Refactoring workflows
a053f2a7 (main) HEAD@{16}: pull --rebase origin main: checkout
a053f2a7c57de1281552356eefad973448fc3379
39813c1f HEAD@{17}: checkout: moving from martin/dev-5586 to ashwinramesh/dev-5599
```

ChatGPT
To undo the latest `git reset` you did (which moved `HEAD` to `HEAD-2`), you'll want to reset the `HEAD` back to the state before that action. In your `git reflog`, the commit before the reset was `3f411213`.

You can reset to this commit using:

```
bash git reset --hard 3f411213
```

This command will move the `HEAD`, index, and working directory to the state of commit `3f411213`. Be cautious with `--hard` as it will discard any changes in your working directory and index that are not committed. If you have any uncommitted changes you want to keep, you might want to use `--soft` or `--mixed` options instead:



See the latest updates to the [Gemini Apps Privacy Hub](#)

You All I have in my pantry is eggs, kraft cheddar and la croix. Help me make something

Show drafts ▾

La Croix won't directly contribute to cooking with your eggs and cheese, but you can definitely make a tasty and satisfying breakfast (or brunch!) with those ingredients. Here are two options:

1. Scrambled Eggs with Cheddar Cheese:

This is a quick and classic dish.

- Crack your eggs into a bowl and whisk them together with a splash of water or milk (not La Croix!).
- Season with salt and pepper.
- Heat a pat of butter or drizzle of oil in a pan over medium heat.
- Pour in the eggs and scramble them until just set.
- Remove from heat and stir in shredded Kraft cheddar cheese.
- You can enjoy this on its own or toast a slice of bread for a base.

er a prompt here

Gemini may display inaccurate info, including about people, so double-check its responses. Your privacy & Gemini

Patterns of Reasoning Can be Induced in LLMs

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

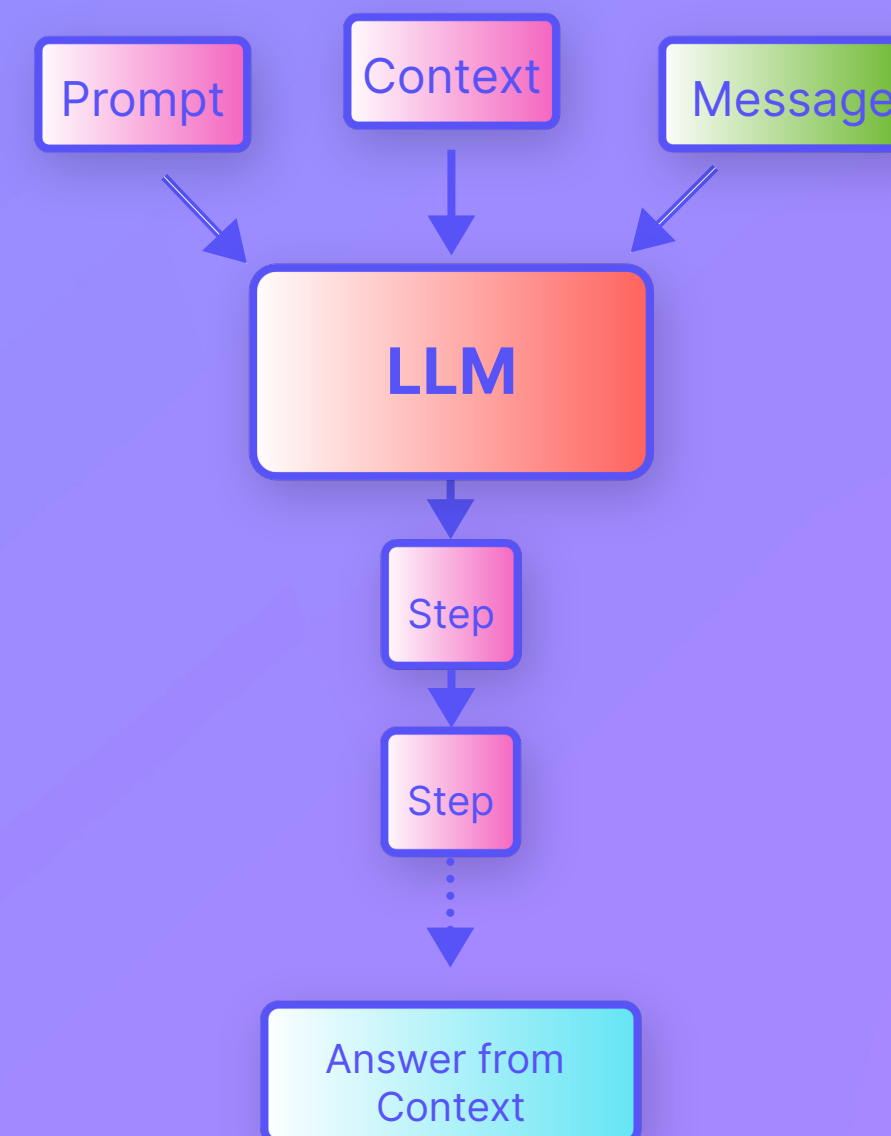
(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓



Things can get pretty wild...

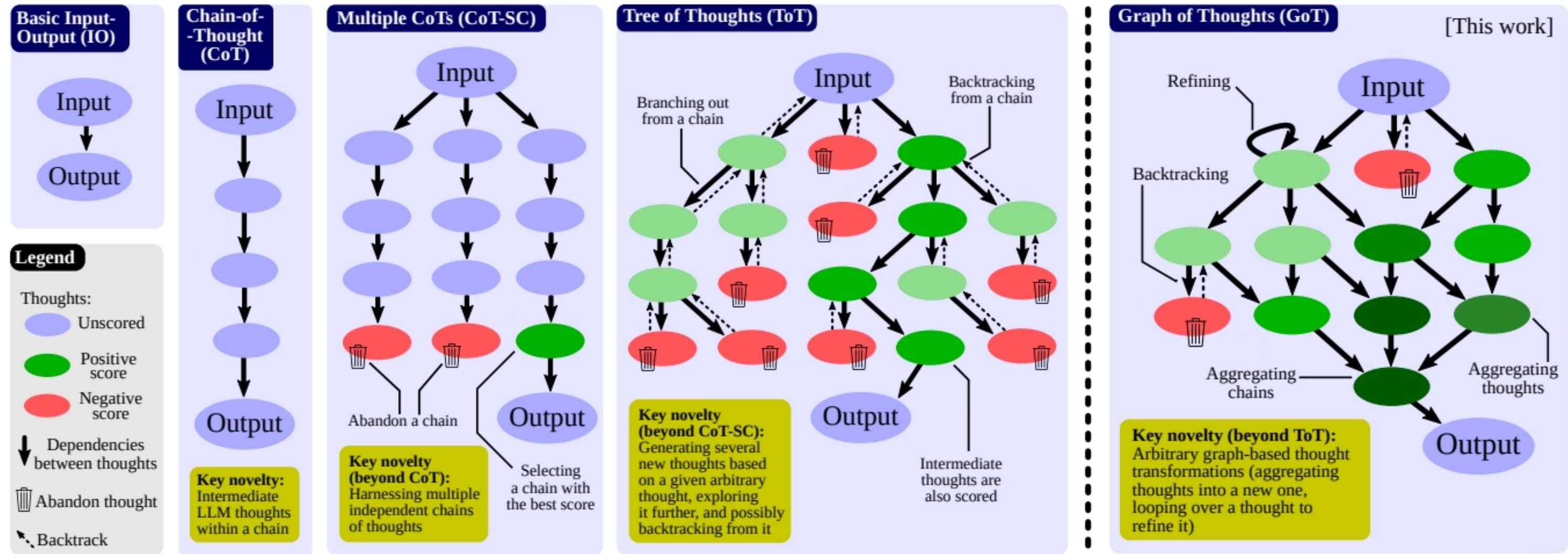


Figure 1: Comparison of Graph of Thoughts (GoT) to other prompting strategies.

LLM Calls are Composable

• Event Loop:

- The event loop is the mechanism that enables Node.js to perform non-blocking I/O operations.
- It continuously checks for and dispatches events or messages in a program.
- It allows Node.js to execute non-blocking operations and schedule callbacks once operations have been completed or events have been received.

In summary, the Node.js programming model is designed to support high-throughput and scalable network applications by leveraging non-blocking I/O and an event-driven architecture. This model enables efficient handling of numerous simultaneous connections with minimal overhead.



You

What's the difference between that and python?



ChatGPT

Node.js and Python are both popular programming environments, but they have distinct differences, especially in their default programming models and execution environments:

• Concurrency Model:

- **Node.js:** Uses an event-driven, non-blocking I/O model. It is designed to perform asynchronous operations, utilizing callbacks, promises, and `async/await` syntax to manage concurrency. The single-threaded event loop handles multiple concurrent operations by offloading I/O operations to the system kernel whenever possible, or using a background

Message ChatGPT...

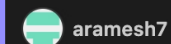
Continual workspace chats aren't used to train our models. ChatGPT can make mistakes.

```
const uniqueMatches = new Map();
uniqueMatches.set(1, 'apple');
uniqueMatches.set(2, 'banana');
uniqueMatches.set(3, 'orange');

const valuesArray = Array.from(uniqueMatches.values());
console.log(valuesArray); // Output: ['apple', 'banana', 'orange']
```

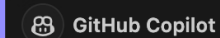
In the example above, we create a `Map` called `uniqueMatches` and add three key-value pairs to it. Then, we use `Array.from()` to convert the values of the `Map` into an array, which is stored in the `valuesArray` variable. Finally, we log the `valuesArray` to the console, which outputs `['apple', 'banana', 'orange']`.

So, in summary, `Array.from(uniqueMatches.values())` is a valid way to retrieve the values from a `Map` and convert them into an array in JavaScript.



aramesh7

For some reason i end up with an empty array



GitHub Copilot

If you're ending up with an empty array, it means that your `Map` object, `uniqueMatches`, doesn't contain any values when you call `Array.from(uniqueMatches.values())`.

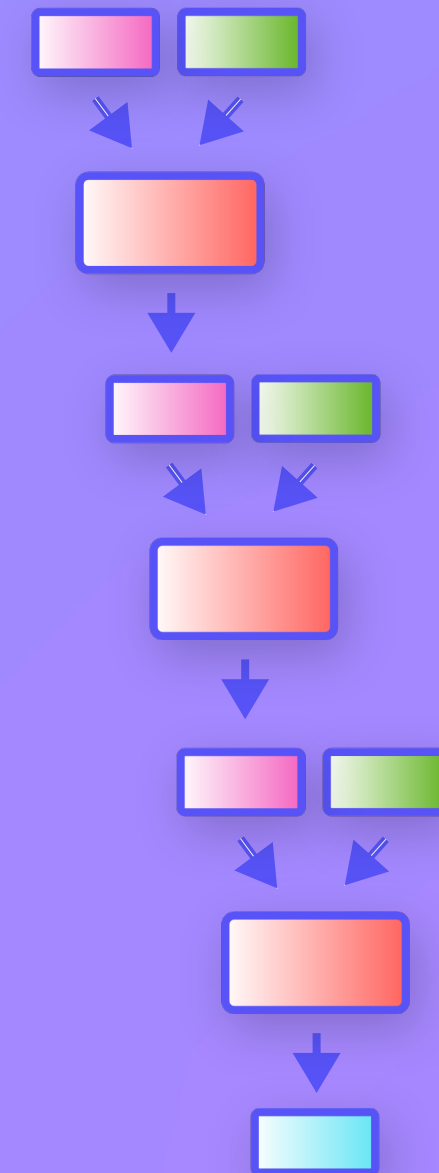
There could be several reasons for this:

1. You might be accessing the values of the `Map` before any values have been set. Make sure that you're calling `Array.from(uniqueMatches.values())` after all the values have been added to the `Map`.
2. If you're adding values to the `Map` asynchronously (for example, in a `Promise` or a `setTimeout`), you might be accessing values before the asynchronous operations have completed. In this case, you need to ensure that all the asynchronous operations have completed before accessing the values.
3. There might be a logic error in the part of your code where you're adding values to the `Map`. Check this part of your code and make sure it's working as expected.

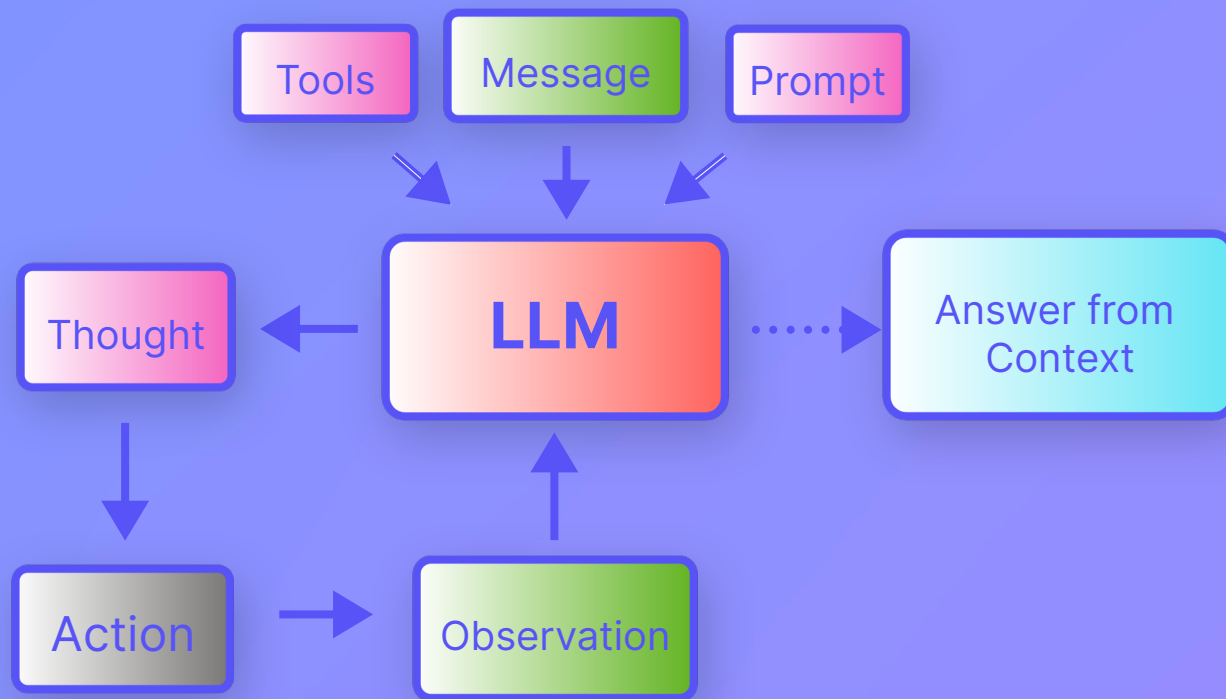
Without seeing the rest of your code, it's hard to say exactly what the issue is. But these are some common reasons why you might be seeing an empty array.

How can I read data from a file in Node.js?

Ask Copilot or type / for commands



LLMs are Agents



```

ReAct

Thought 1: I need to search Apple Remote and find the program
it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in
October 2005 by Apple ... originally designed to control the
Front Row media center program ...

Thought 2: Apple Remote was originally designed to control the
Front Row media center program. I need to search Front Row
next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat
to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search Front Row
(software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software ...

Thought 4: Front Row (software) is controlled by an Apple
Remote or the keyboard function keys. So the answer is
keyboard function keys.
Act 4: Finish[keyboard function keys]
    
```

```

1. Decision making

(a) Task
You are in the middle of a room
[...] Task: clean some pan and put
it in countertop.

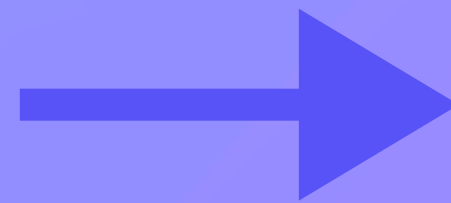
(b) Trajectory
[...]
Action:take pan1 from stoveburner1
Obs:Nothing happens. [...]
Action:clean pan1 with sinkbasin1
Obs:Nothing happens. [...]

(c) Evaluation
Rule/LM Heuristic:
Hallucination.
(internal / external)

(d) Reflection
[...] tried to pick up the pan in
stoveburner 1 [...] but the pan
was not in stoveburner 1. [...]

(e) Next Trajectory
[...] Action: take pan 1 from
stoveburner 2
[...] Obs: You put the pan 1 in
countertop 1.
    
```


LLMs 🤝 APIs



```
const response = await openai.chat.completions.create({  
  model: "gpt-3.5-turbo",  
  messages: messages,  
  tools: tools,  
  tool_choice: "auto",  
});
```

LLMs 🤝 APIs

You



SHOP NAME
Address: Lorem Ipsum, 23-10
Telp. 11223344

CASH RECEIPT

Description	Price
Lorem	1.1
Ipsum	2.2
Dolor sit amet	3.3
Consectetur	4.4
Adipiscing elit	5.5

Total 16.5
Cash 20.0
Change 3.5

Bank card # 234
Approval Code #123456

THANK YOU!

designed by freepik

Could you tally the items on this and get me the average price of the items?

ChatGPT

Message ChatGPT...

ChatGPT

Finished analyzing

Always expand output?

```
python  
  
from PIL import Image  
import pytesseract  
  
# Open the image file  
img_path = '/mnt/data/image.png'  
img = Image.open(img_path)  
  
# Use tesseract to do OCR on the image  
text = pytesseract.image_to_string(img)  
  
# Output the text content to check what was recognized.  
text
```

Result

```
'SHOP NAME\n\nAddress: Lorem Ipsum, 23-10\nTelp. 11223344\n\nCASH RECEIPT\n\n'
```

Finished analyzing

Always expand output?

```
python  
  
# Extracting individual prices from the recognized text to calculate the average price
```

Message ChatGPT...

Result

```
([11223344.0, 14.0, 2.2, 3.3, 4.4, 5.5, 16.5, 20.0, 35.0, 234.0],  
1122367.8900000001)
```

The OCR recognized the following list of numbers as prices:

- 11,223,344 (which appears to be a misinterpretation of the telephone number as a price)
- 14
- 2.2
- 3.3
- 4.4
- 5.5
- 16.5
- 20 (this is the cash given, not an item price)
- 3.5 (this is the change given, not an item price)
- 234 (which appears to be part of the bank card number, not a price)

To get the average item price correctly, we need ●

Message ChatGPT...

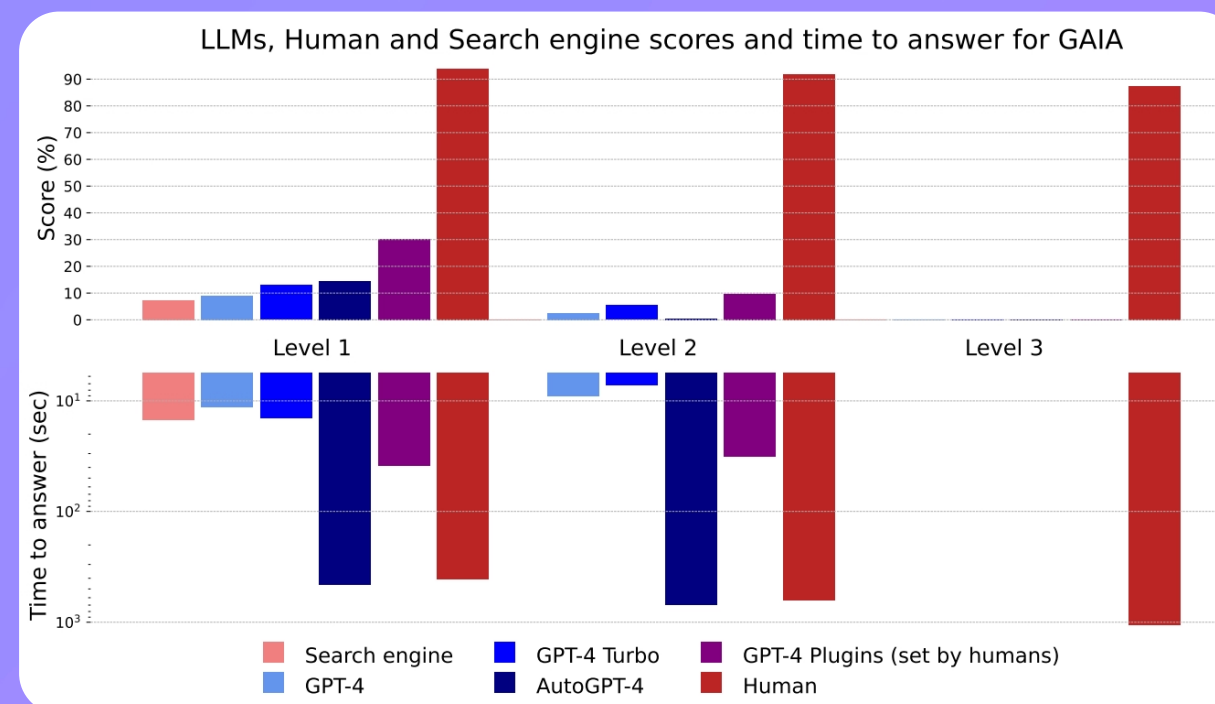
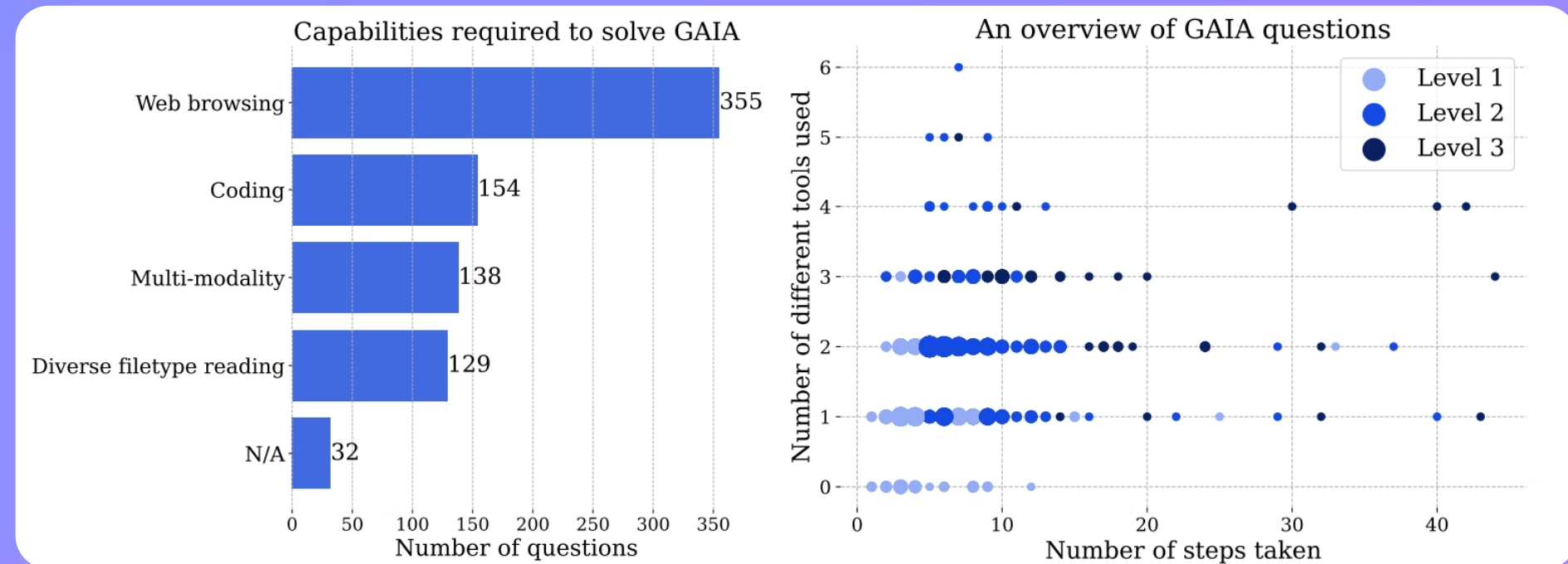


BUT IS IT AGI?



GAIA

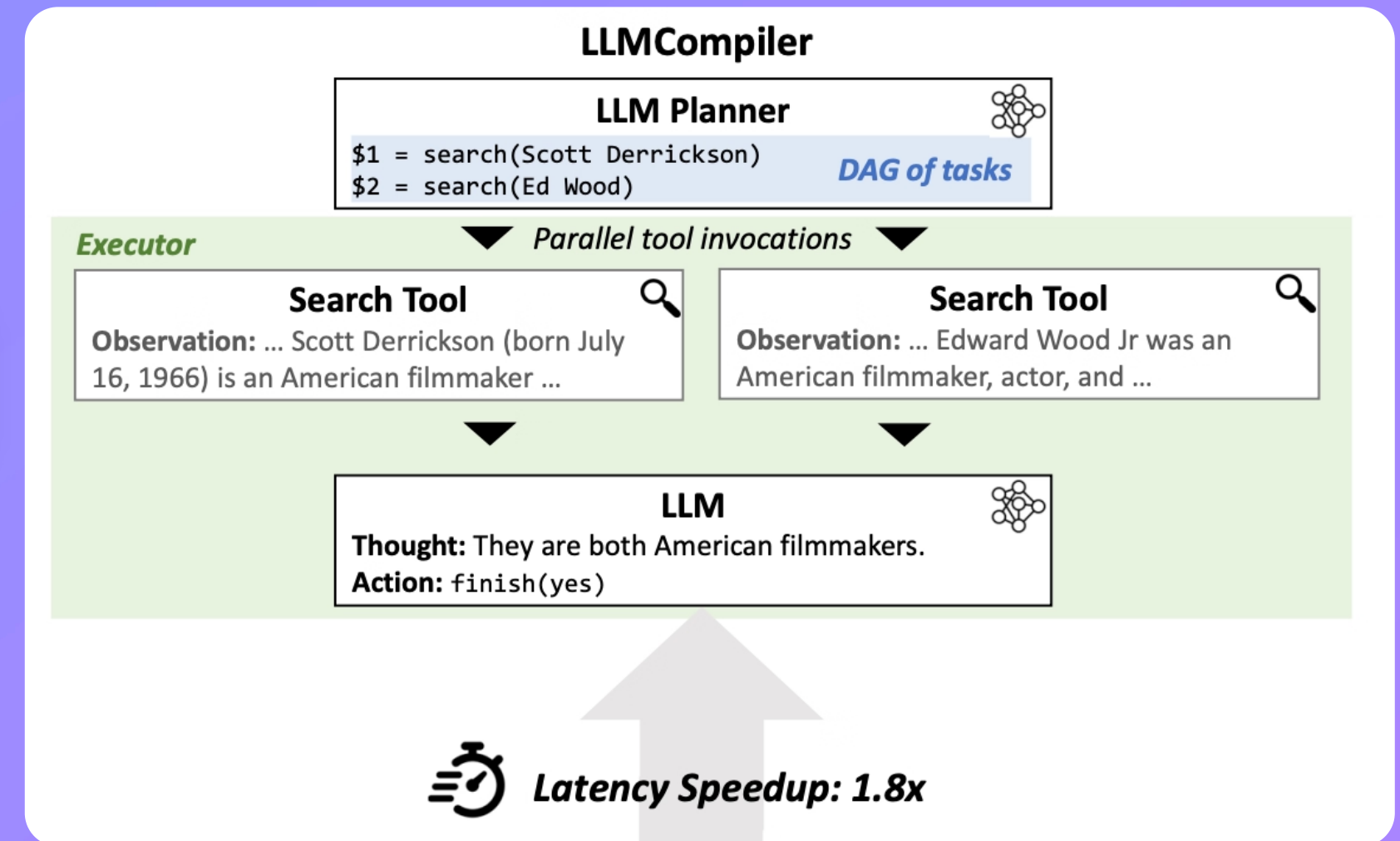
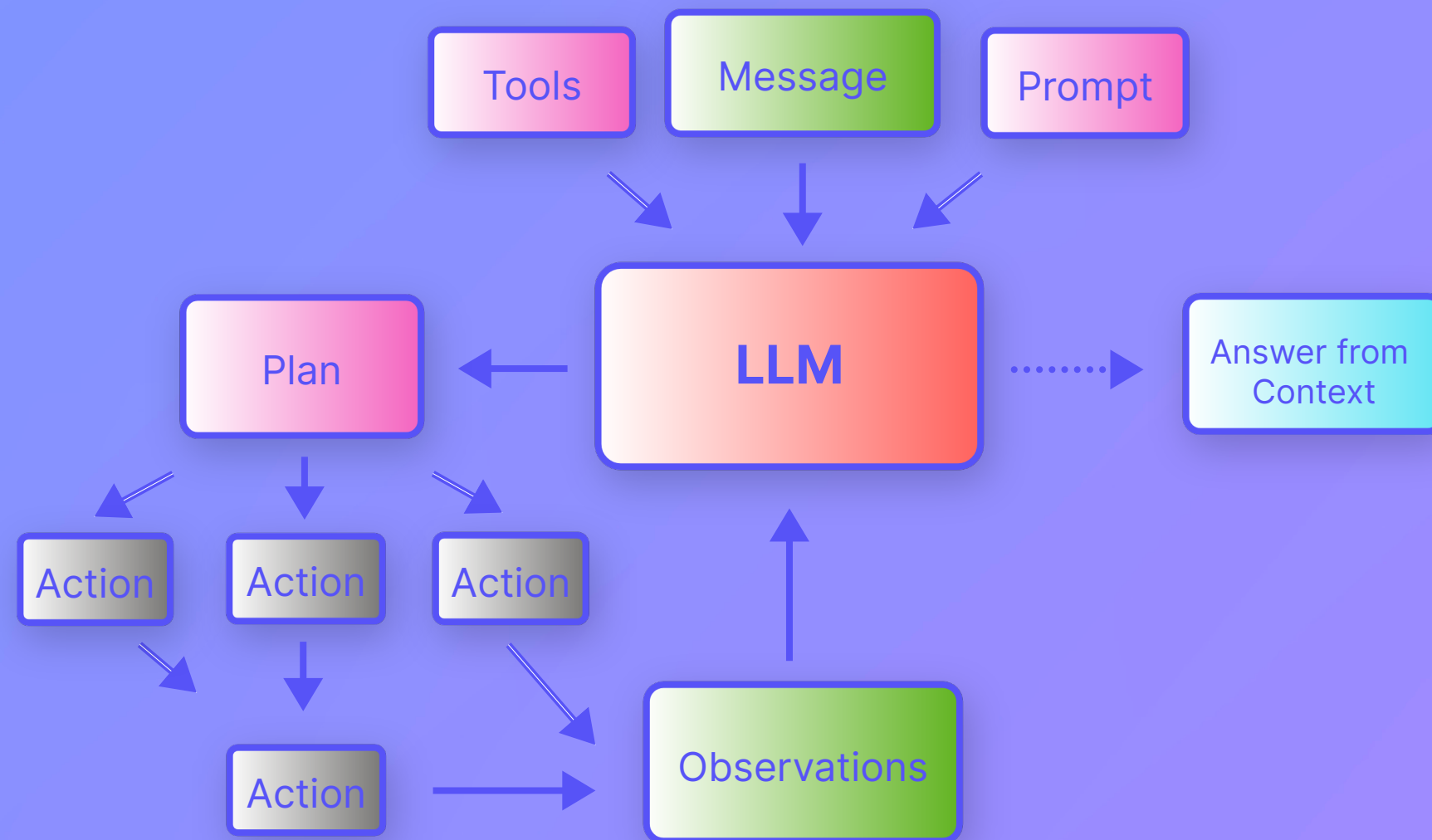
Question <i>string · lengths</i>	Level <i>string · classes</i>	Final answer <i>string · lengths</i>
On June 6, 2023, an article by Carolyn Collins Petersen was published in Universe Today. This article mentions a team that produced a paper about their observations, linked at the bottom of the article. Find this paper. Under what NASA award number was the work performed by R. G. Arendt supported by?	1	80GSFC21M0002
The attached file lists the locomotives owned by a local railroad museum. It gives each locomotive's identifying number, operating status, and the name of the daily excursion it heads, if operational. What are the odds that today's Sunset Picnic Trip will use a steam locomotive? Assume that each day's excursion picks one of its assigned locomotives at random, and express the answer in the form "1 in 4", "1 in 5", etc.	2	1 in 3
The year is 2022. I am at the National Air and Space Museum east of the Potomac River. I want to go to Fire Station 301 DCA ARFF using the metro. I go in the wrong direction and end up at the station closest to Cleveland Elementary School. How many metro stations am I away from my original destination if I don't change lines? Your answer should be a numerical integer value.	3	8



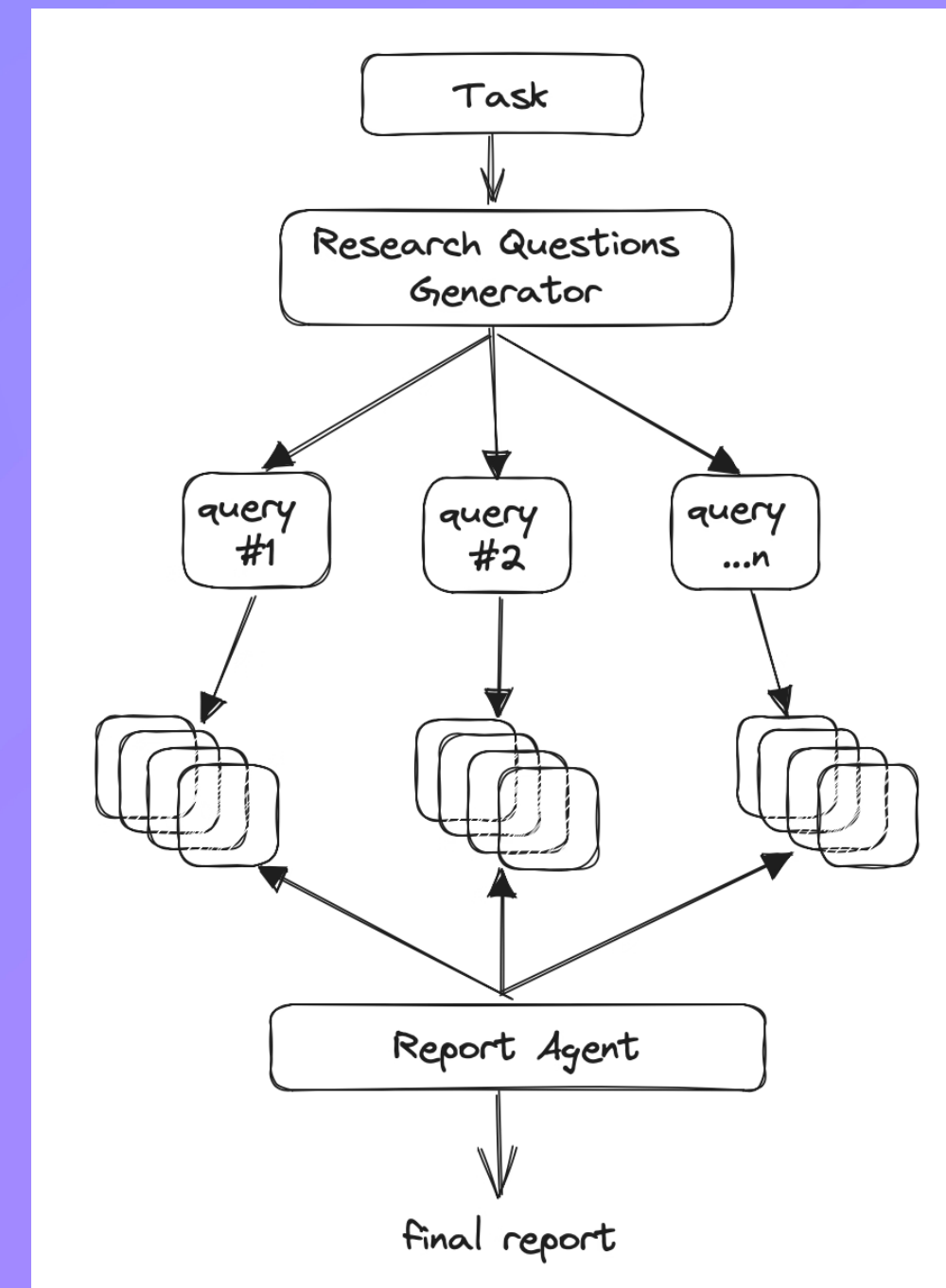
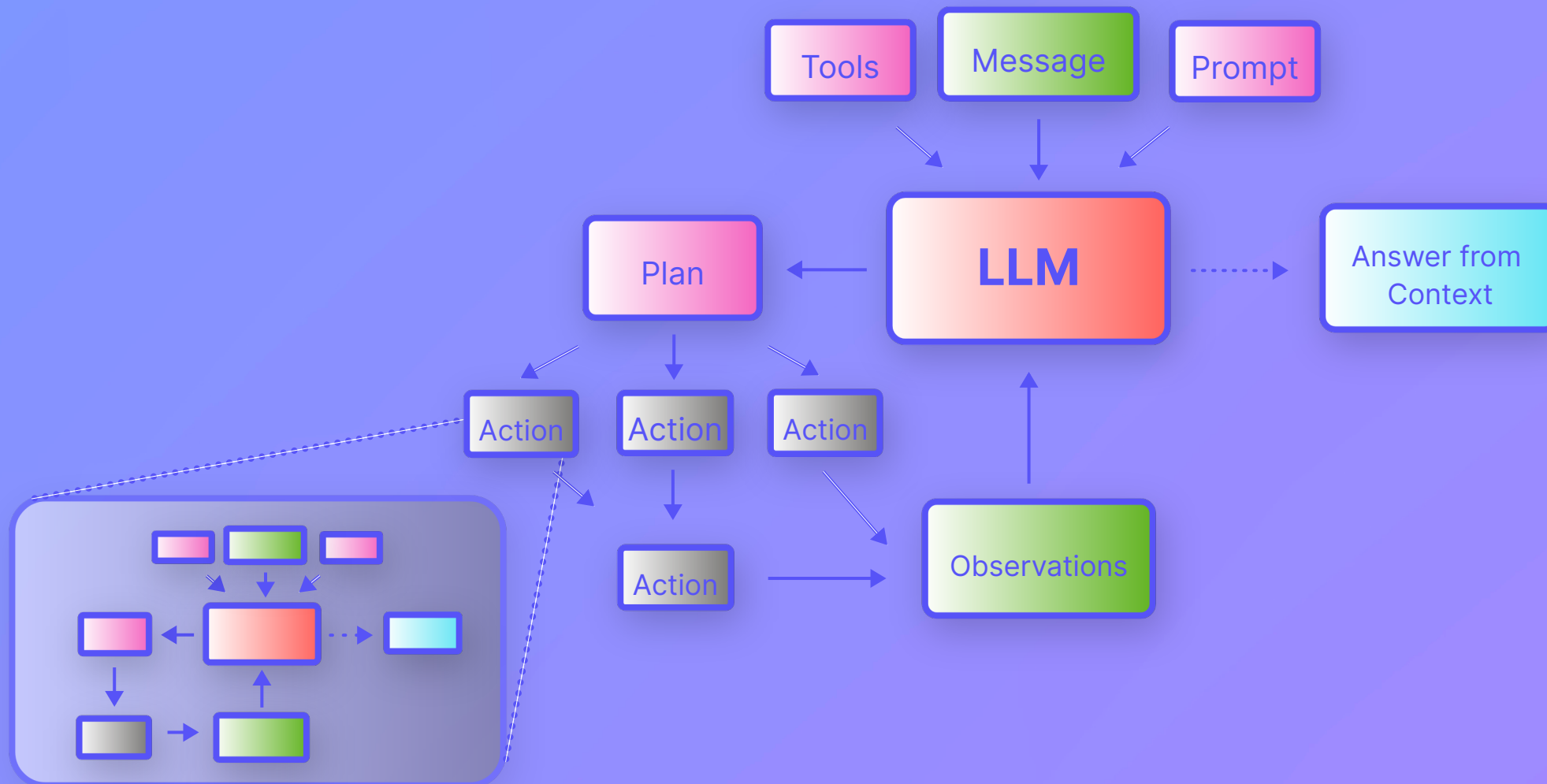
General AI Assistants

- Pursue long term plans and goals
- Learn from mistakes and develop skills
- Build and use new tools

Planning



Delegation



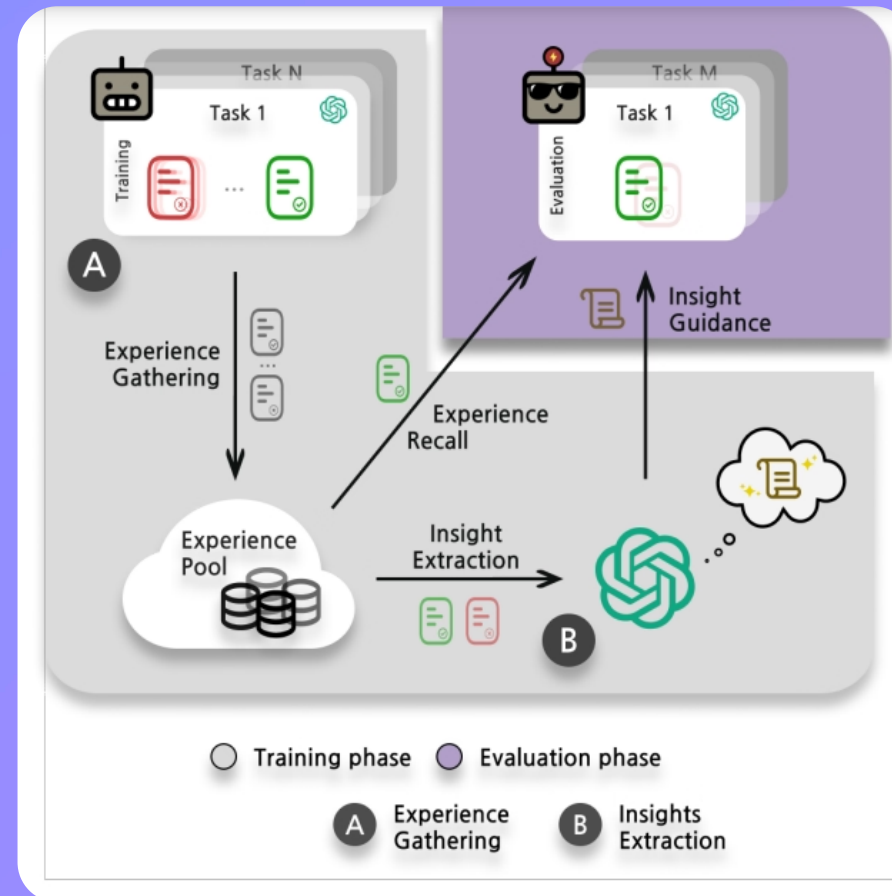
Memory

Manage Memory

ChatGPT

- Has a 2 year old daughter named Lina
- Daughter, Lina, loves jellyfish
- Prefers meeting summaries to have headlines with bullets and action items summarized at the end.
- Prefers assistance with writing blog posts to be more concise, straightforward, and less emotive.
- Loves to travel.
- Is interested in traveling to Mexico for April vacation.

Clear ChatGPT's memory



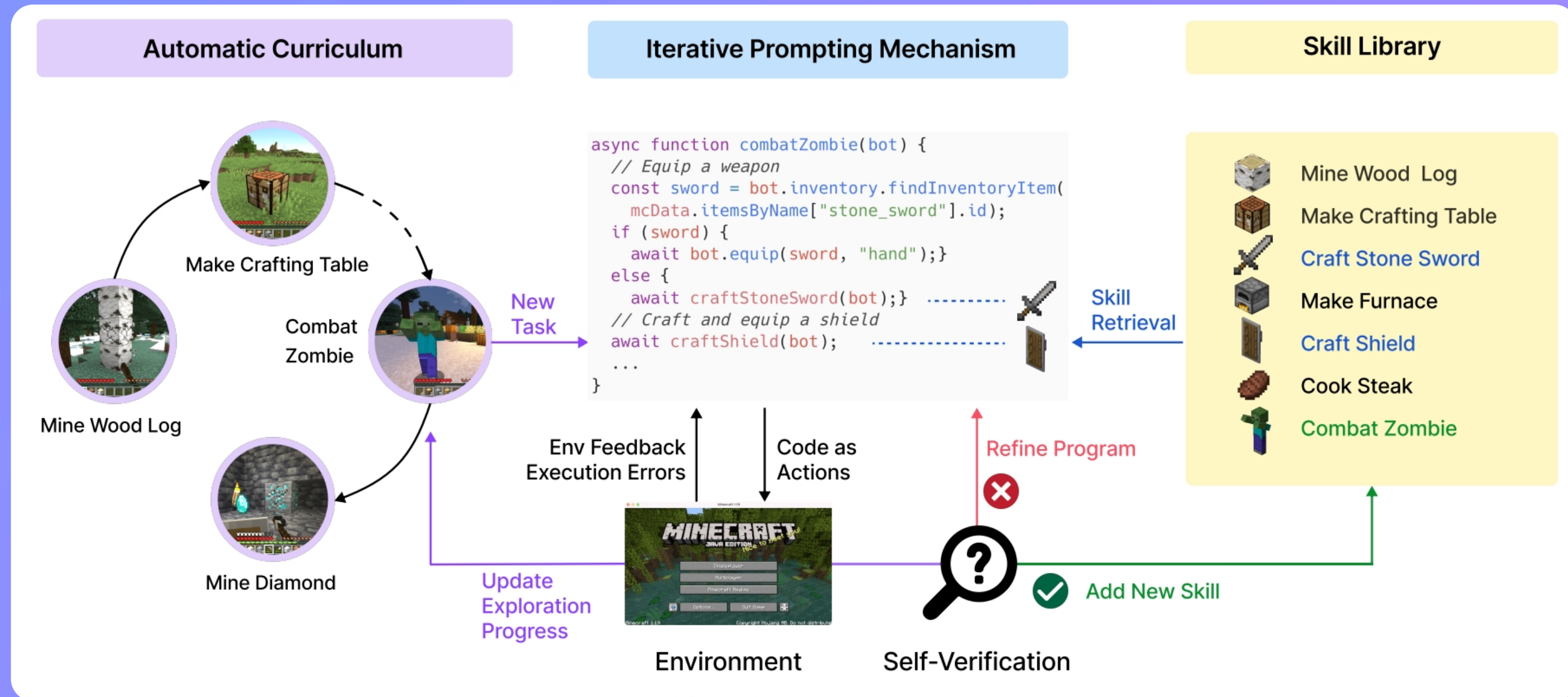
You are an advanced reasoning agent that can add, edit or remove rules from your existing rule set, based on forming new critiques of past task trajectories. You will be given...

<p>two previous task trials in which you ...</p> <p>[Task description] ...</p> <p>one successful and one unsuccessful trial. You failed the trial because...</p> <p>[Task failure reasons].</p> <p>Here are the two previous trials to compare and critique:</p> <p>[Failed/Succeeded Trajectories]</p>	<p>successful tasks trials in which you ...</p> <p>[Task description].</p> <p>Here are the trials:</p> <p>[Succeeded Trajectories]</p>
--	--

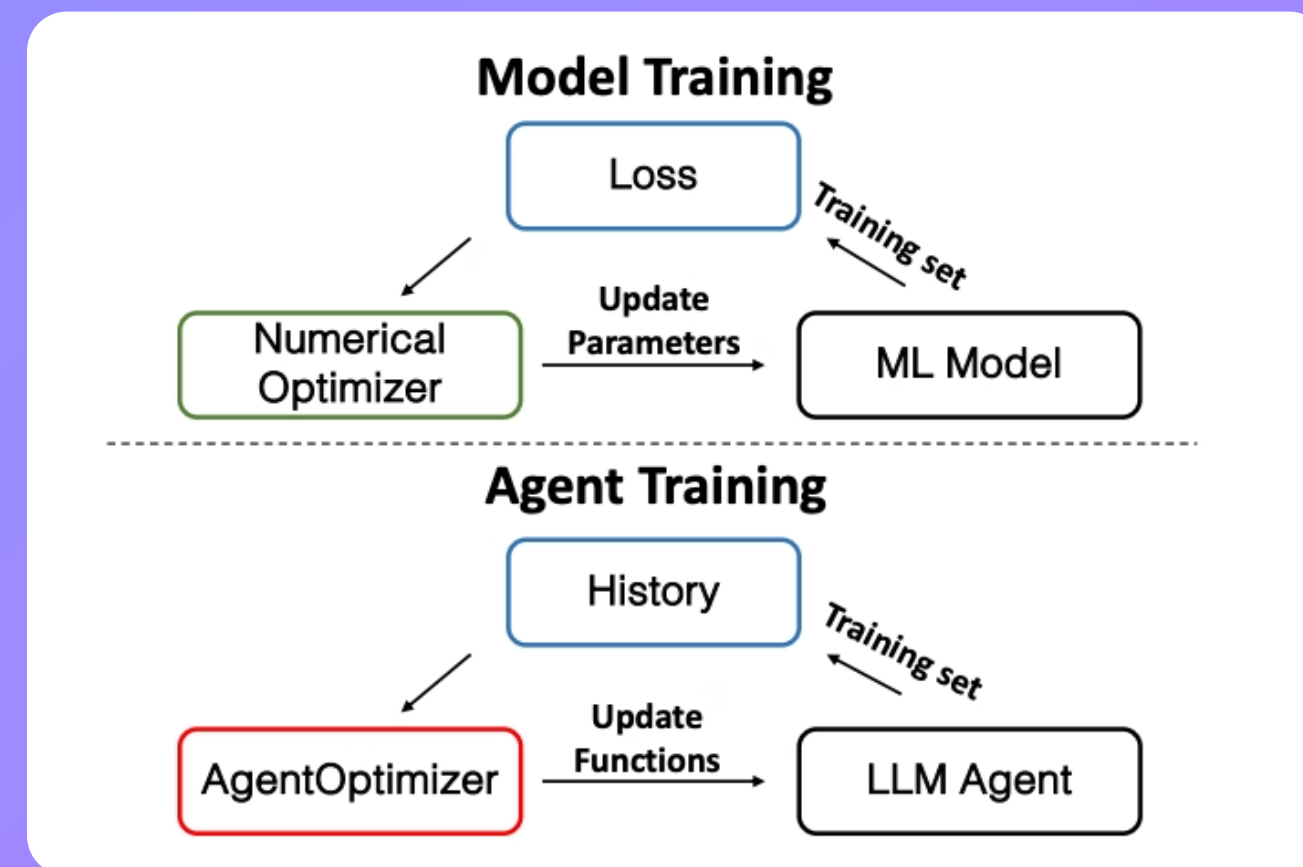
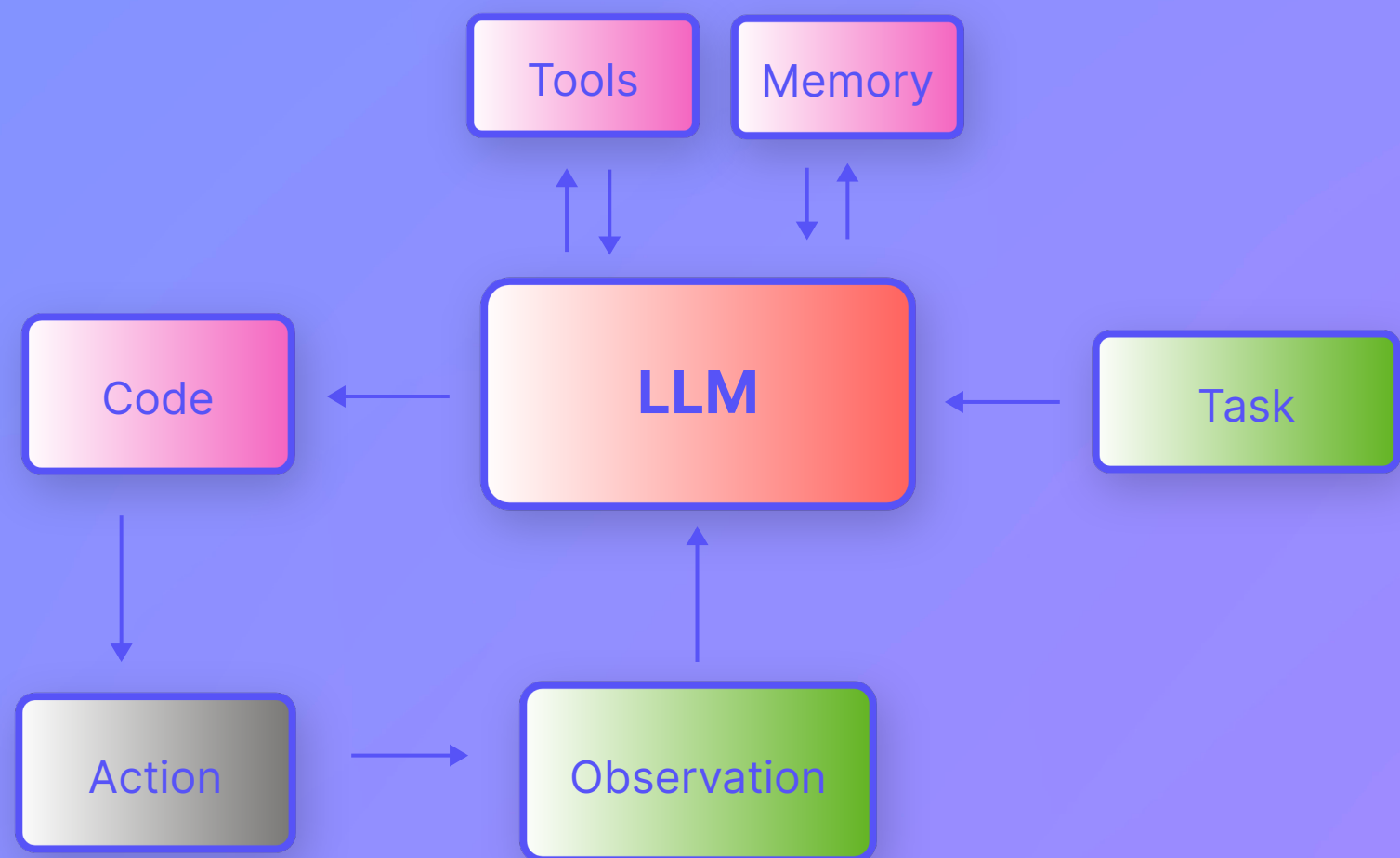
Here are the EXISTING RULES:

[Currently existing insights]

Putting It Together



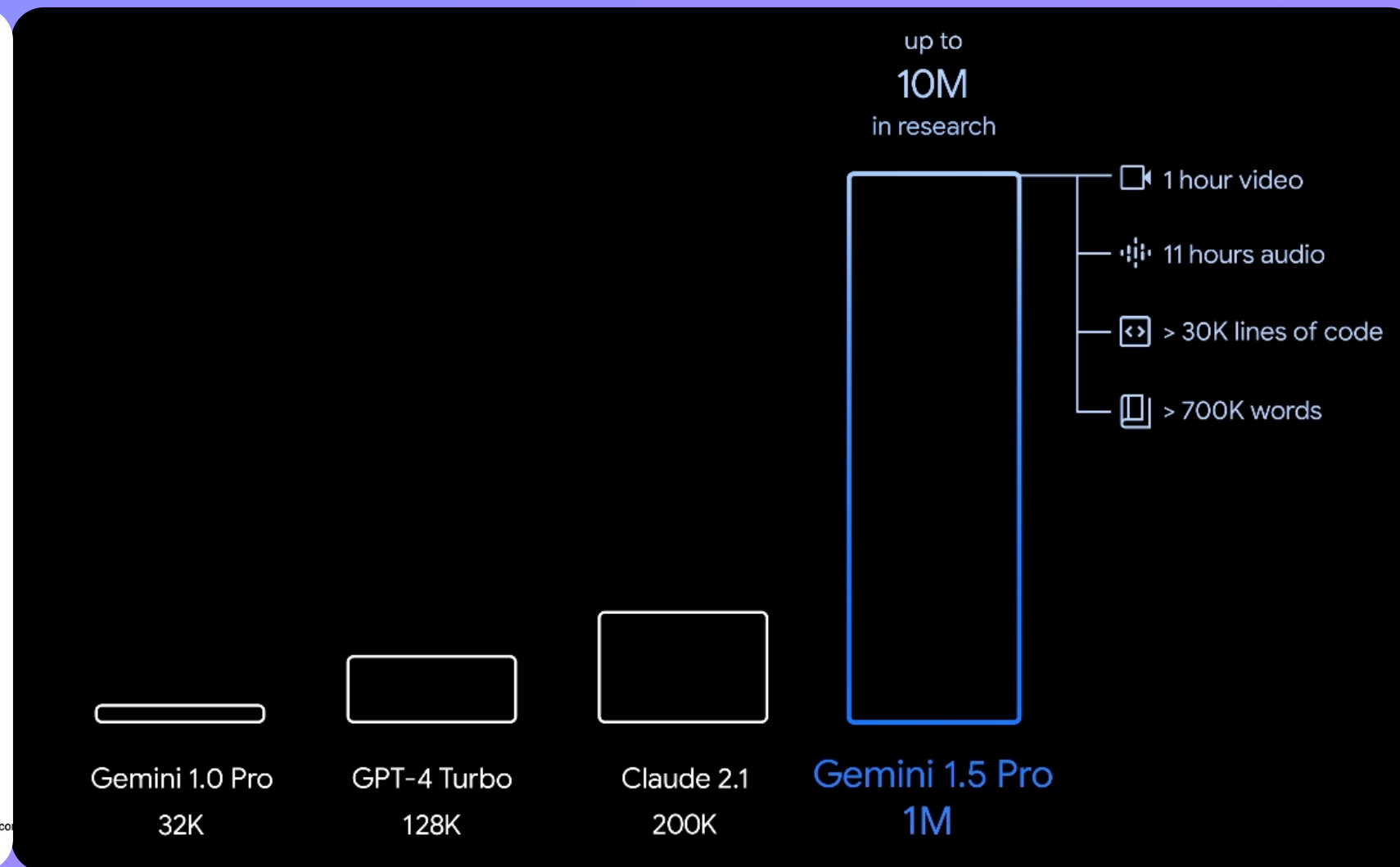
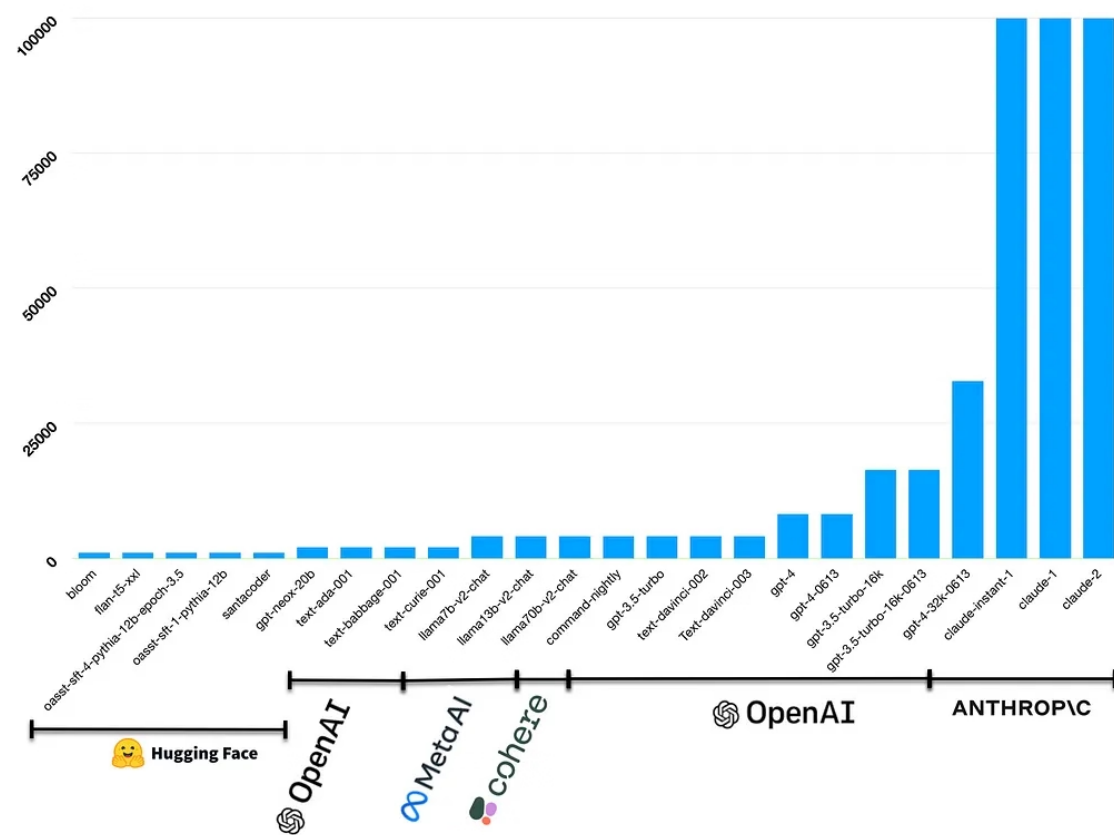
Training General Assistants



Isn't all this really slow?

Constraints are Lifting

Large Language Model Context Size



Constraints are Lifting

groqcloud

Playground

USER: I have eggs, bread and cheese. Whats the most complicated dish I can make?

ASSISTANT: With eggs, bread, and cheese, you have the ingredients for a variety of dishes. Here are a few ideas for more complicated dishes you can make:

1. Eggs Benedict: Toast the bread, top it with a slice of cheese, a poached egg, and some ham or Canadian bacon.
2. Eggs Florentine: Similar to Eggs Benedict, but substitute spinach for the ham.
3. Eggs Sardou: Similar to Eggs Benedict, but add a layer of artichoke hearts and hollandaise sauce.
4. French toast: Soak bread slices in a mixture of eggs, milk, and cinnamon, then fry them in a pan until golden brown. Serve with maple syrup, or fresh fruit.
5. Cheese soufflé: Beat eggs with cheese, milk, and seasonings, pour into a baked soufflé dish, and bake until puffed and golden.
6. Quiche Lorraine: Mix eggs, cheese, and milk with diced ham or bacon, pour into a pie crust, and bake until set.
7. Strata: Layer bread, cheese, and diced ham or bacon in a baking dish, pour eggs over the top, and bake until puffed and golden.
8. Egg and cheese tart: Mix eggs, cheese, and diced ham or bacon, pour into a pre-baked tart crust, and bake until set.

These are just a few ideas to get you started. With eggs, bread, and cheese, the possibilities are endless!

+ New Message

Round Trip Time: 1.90 s	Model: llama2-70b-4096		
Speed (T/s)	Input	Output	Total
	1920	291	322
Tokens	48	377	425
Inference Time (s)	0.03	1.3	1.32

Inference: 1321 ms ⚡ Tokens/s: 291

View code

Summary

- Move towards general AI assistants! Forget prompting, use frameworks.
- Target an agent architecture that supports
 - Planning over long horizons
 - Continual learning and knowledge acquisition
 - Tool creation
- Build a gym to do online training
 - Generate synthetic data or use a simulator
 - Optimize higher level components like tools, prompts, and memory rather than model parameters

Summary

- Move towards general AI assistants! Forget prompting, use frameworks.
- Target an agent architecture that supports
 - Planning over long horizons
 - Continual learning and knowledge acquisition
 - Tool creation
- Build a gym to do online training
 - Generate synthetic data or use a simulator
 - Optimize higher level components like tools, prompts, and memory rather than model parameters
- Until GPT5 Comes out...

Thank you!