CSCE 689: Special Topics in Advanced NLP

Lecture 19: Model Memory, Agents

Kuan-Hao Huang

Fall 2026



Course Project – Final Report

- Due: 12/6
- Page limit: 7 pages (excluding reference)
- The report should include
  - Introduction to the topic you choose and problem definition
  - Related literature and overview of existing progress and challenges
  - Proposed solutions, novelty, and contributions
  - Implementation details
  - Experimental settings, including dataset, models, baseline descriptions, evaluation metrics
  - Results, findings, and conclusion
- Make it a complete report!

Final Presentation

- Each team has 7 minutes for presentation
  - You have to stop once you reach 7 minutes
- The presentation should include
  - The topic you choose
  - Novelty/challenges
  - Your approach
  - Experiments/evaluation
  - Results, findings, insights
- Clarity is important
  - Teach your classmate about your topic
- Time management is also important

Presentation Order

- Final presentation is online

11/30

1. Faaiz Umer , Oseluolemen Inotu
2. Yiming Xiao, Shuning Gu
3. Mingyang Wu, Siyuan Yang
4. Chedi Morchdi
5. Jason Kim, Andrew Vong
6. Akshay Hitesh Shah, Saisugash Senthil
7. Christopher Kim, Yihong Yang
8. Ali Taheri, Damandeep Singh
9. Amir Suhail, Fredy Cortez

12/2

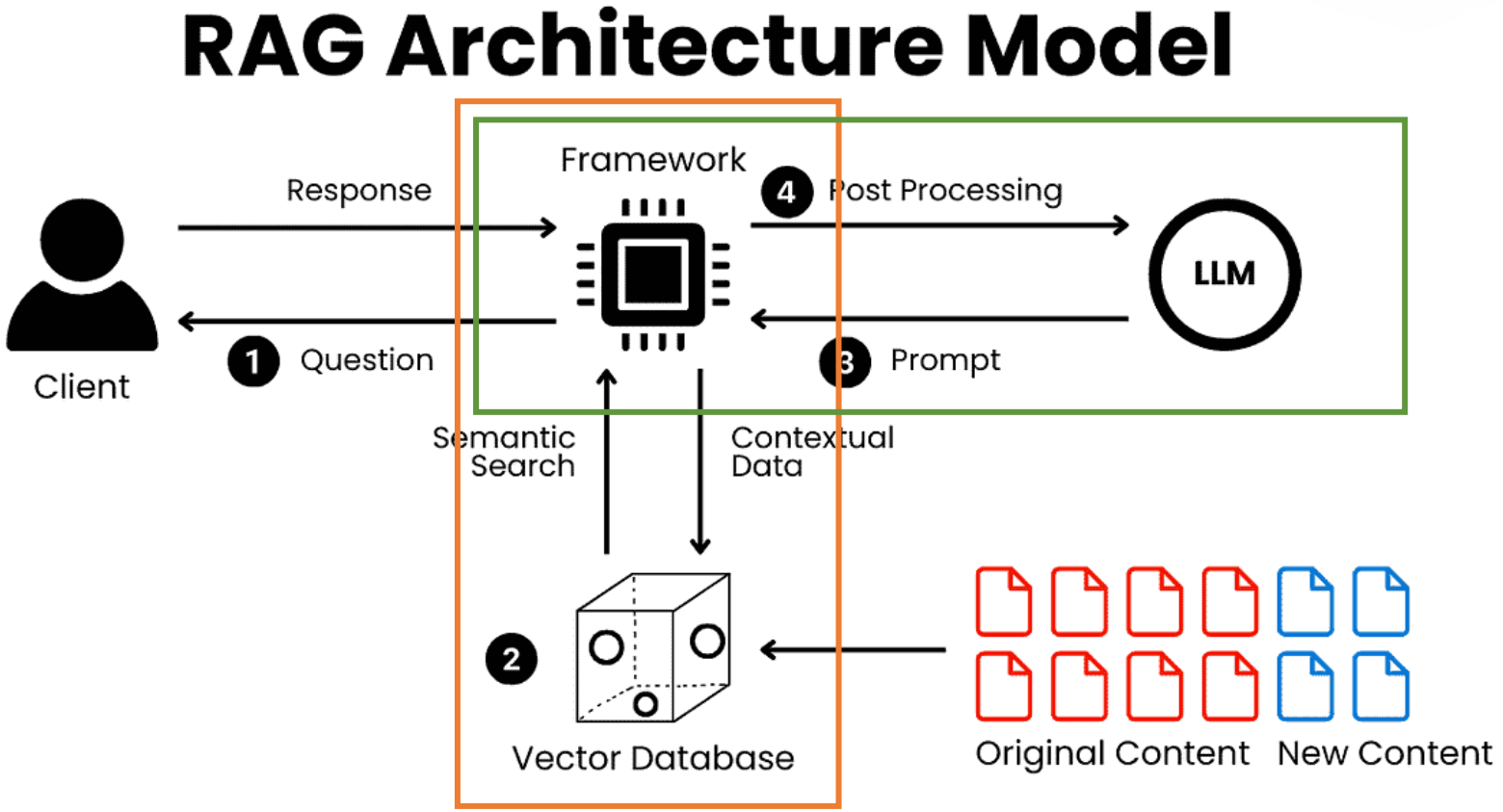
1. Vaishnavi Poti, Ed Stone
2. Taeyoon Kwon, Ming Yan
3. Karan Pramod Sharma, Mrunalini Thamankar
4. Han-Ju Chen
5. Lawrence Wong, Omer Burak Cinar
6. Raphael Zhu, Ruei-Chi Chen
7. Chenghao Qiu
8. Prithvi Patel

Invited Talks

- **Date:** 11/23 online @ Zoom
  - https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1
  - **Talk:** TBD
  - **Speaker:** Da Yin, Founding Member of Technical Staff at NeoCognition

Recap: Retrieval-Augmented Generation (RAG)

- Information Retrieval + Large Language Models

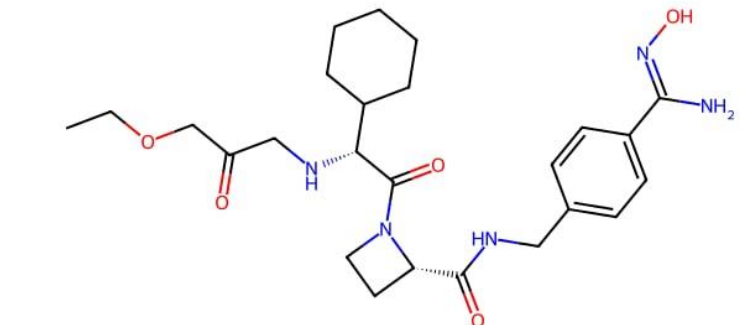


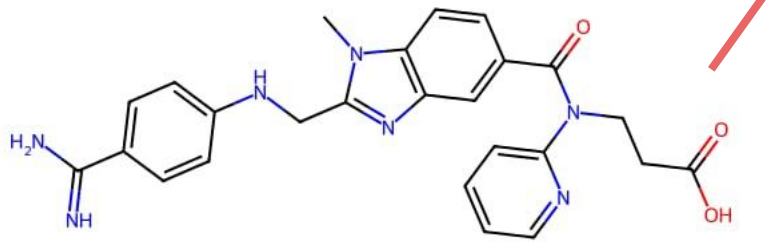
https://blog.gopenai.com/langchain-and-its-role-in-building-robust-rag-systems-2808b7e5ddb9

Tool-Augmented Language Models

- We can treat search engine as an external tool for LLMs
  - Can we let LLMs use more tools?

LLMs Lack of Domain Knowledge





what is the name of this molecule:
CCOCC(=O)CN[C@@H]
(C(=O)N1CC[C@H]1C(=O)NCc1ccc(/C(N)=N/O)cc1)C1CCCCC1

The SMILES you gave:

```
CCOCC(=O)CN[C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(/C(N)=N/O)cc1)C1CCCCC1
```

corresponds to the **direct thrombin inhibitor drug**:

Dabigatran

More precisely, the structure matches **dabigatran (the active drug)** rather than the orally used prodrug **Dabigatran etexilate**.

Key structural features that identify it

From the SMILES:

- **Amidoxime group**: `/C(N)=N/O`
  - characteristic benzamidine bioisostere used in thrombin inhibitors
- **p-substituted benzyl amide**
- **proline-like tertiary amide ring**: `N1CC[C@H]1`
- **cyclohexyl substituent**: `C1CCCCC1`

Give LLMs a database

LLMs Don’t Know News

What is 2026 WBC US vs Canada result?

Assistant B

There is **no result** yet for a 2026 WBC game between the USA and Canada because **the tournament has not happened yet.**

The 2026 World Baseball Classic is scheduled to take place in **March 2026**.

Here is what we know so far:

- **Status:** The brackets and schedule have not been announced.
- **Qualification:** Both Team USA and Team Canada automatically qualified for the 2026 tournament based on their performance in the 2023 WBC.
- **Likelihood:** While they are often placed in the same pool due to geography, it is not guaranteed they will play each other in the group stage until the official schedule is released (likely in late 2025).

Team USA beats Canada to set up WBC semis showdown vs. D.R.



Jeff Passan
Mar 13, 2026, 11:21 PM ET

Follow

Share Like 3.3K

HOUSTON -- In the aftermath of the Dominican Republic's mercy-rule win over Korea to advance to the World Baseball Classic semifinals, Dominican fans at LoanDepot Park in Miami made clear what they wanted to see next, chanting: "Queremos a USA" -- we want the USA.

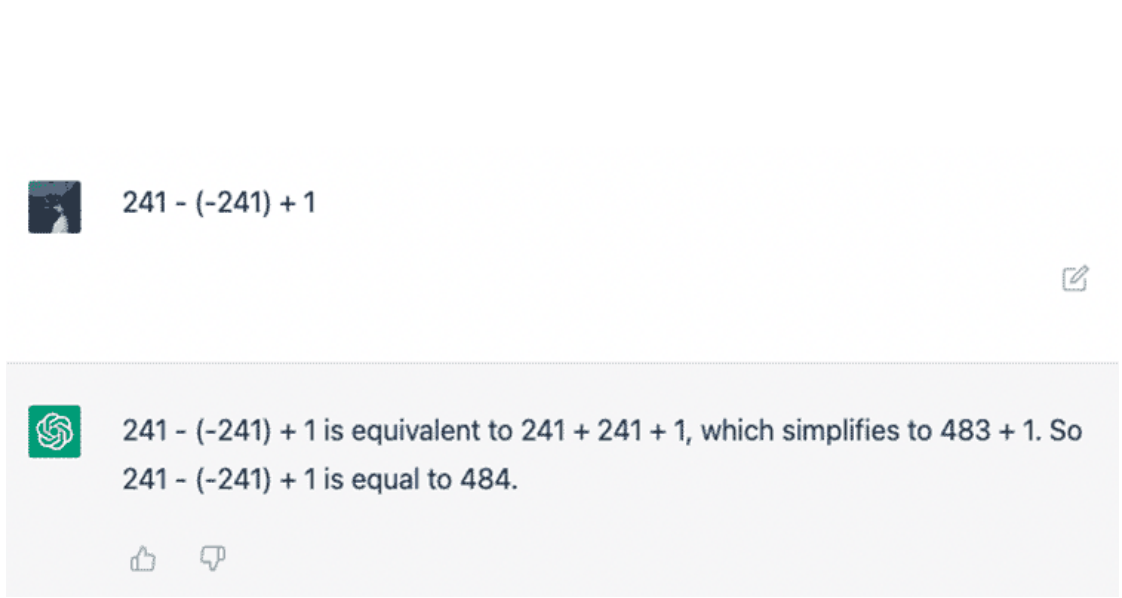
They got what they asked for.

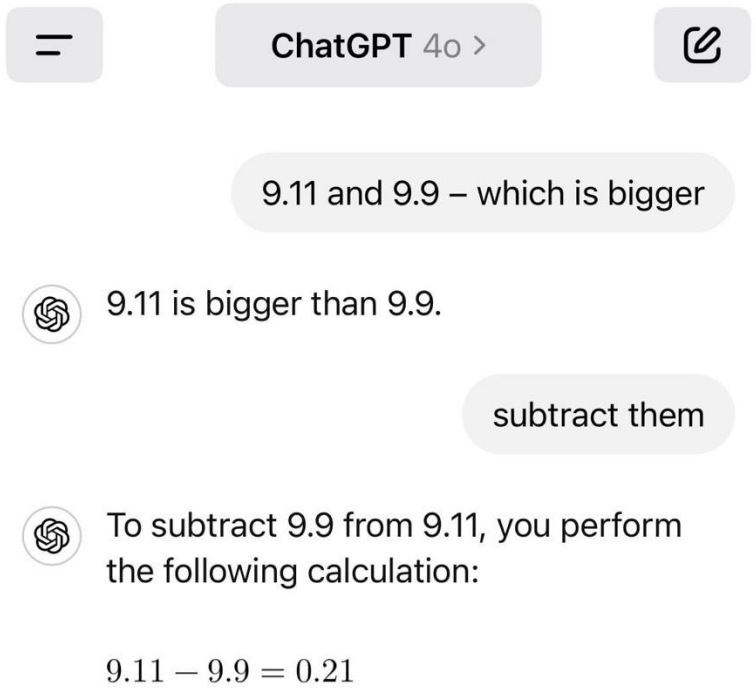
The United States booked its spot in the WBC semifinals with a 5-3 victory against Canada on Friday night, setting up a dream clash between the two most talented teams in the tournament. National League Cy Young winner Paul Skenes will start for Team USA against the formidable Dominican lineup that has wrecked opposing pitchers, while Athletics right-hander Luis Severino is expected to face an American lineup big on names but not so much on production thus far.

Give LLMs a search API

https://arena.ai/
https://www.espn.com/mlb/story/_/id/48201228/team-usa-beats-canada-set-world-baseball-classic-semifinal-showdown-vs-dominican-republic

LLMs are Bad at Math





Give LLMs a calculator

https://www.reddit.com/r/singularity/comments/1e4fcxm/none_of_current_llms_can_truly_reason_and_cannot/
https://telefonicatech.com/en/blog/llm-mathematical-challenges

LLMs Generate Code with Errors

Problem Description: *APPS-test-673*

```
Find the smallest integer x greater than n,
so it is divisible by the number k.
The input only line contains two integers
n and k (1 ≤ n, k ≤ 10^9).
-----Examples-----
Input
5 3
Output
6
```

example test case

GPT3 Output:

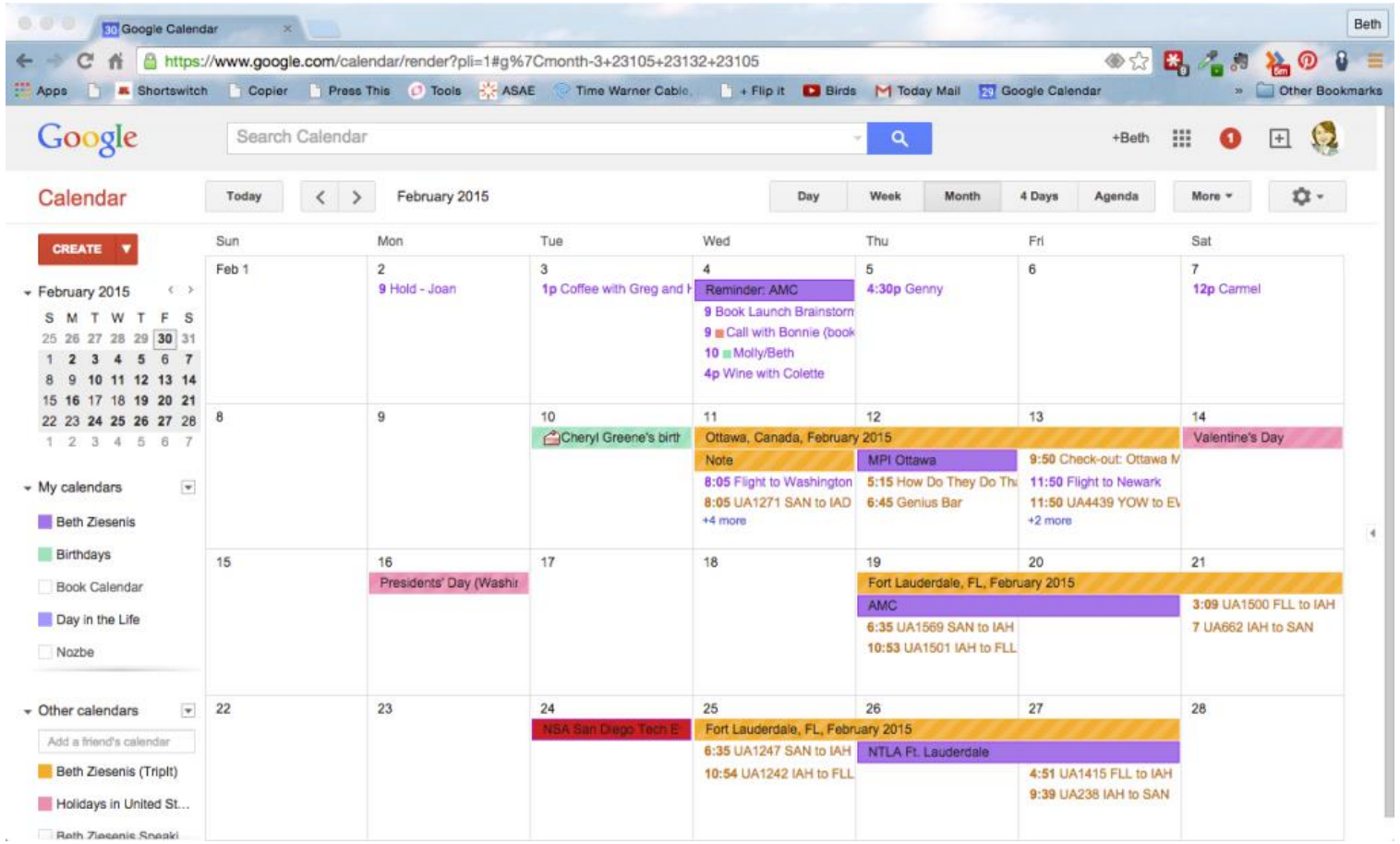
```
n, k = map(int, input().split())
if n % k == 0:
      print n + k
else:
      print n + k - (n % k)
```

Error Message:

SyntaxError: Missing parentheses in call to 'print'. Did you mean print(n + k)?

Give LLMs a Python interpreter

LLMs Don’t Know Dates



Give LLMs a calendar API

https://yournerdybestfriend.com/2017/03/31/google-calendar-the-calendar-to-rule-them-all/

Other Tools

- Translator
- Automatic Speech Recognition
- PDF Parser
- Object Detection
- Optical Character Recognition
- Text-to-Speech
- …

Tool-Augmented Language Models

- LLMs don’t need to be experts in everything
- LLMs act as coordinators that know when to consult experts.

Tool-Augmented Language Models

The New England Journal of Medicine is a registered trademark of **[QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society]** the MMS.

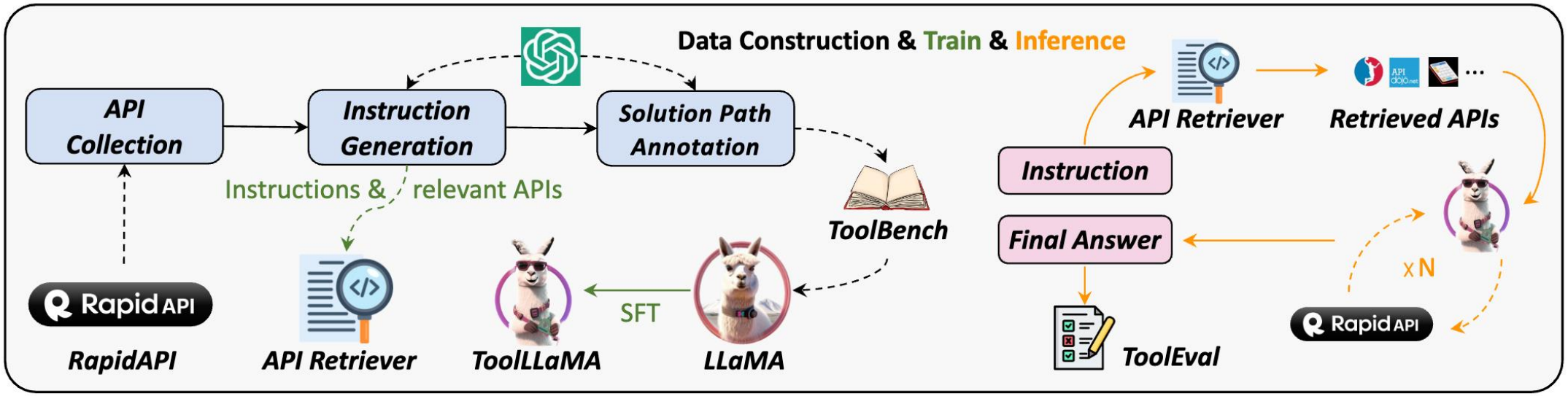
Out of 1400 participants, 400 (or **[Calculator(400 / 1400) → 0.29]** 29%) passed the test.

The name derives from "la tortuga", the Spanish word for **[MT("tortuga") → turtle]** turtle.

The Brown Act is California's law **[WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.]** that requires legislative bodies, like city councils, to hold their meetings open to the public.

Toolformer: Language Models Can Teach Themselves to Use Tools, 2023

Training Tool-Augmented Language Models



ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs, 2023

Tool Meta Data

Image Captioner Tool

tool_description="A tool that generates captions for images."

inputs={image, prompt}

outputs={str: "The generated caption."}

demos=tool.execute(image="image.png", prompt="Describe the image.")

user_metadata={"limitation": "May make mistakes in complex scenes & counting."}

Object Detector Tool

tool_description="A tool that detects and saves objects in an image"

inputs={image,labels,threshold}

outputs={list: objects, scores, boxes}

demos=tool.execute(image="car.png", labels=["car"])

user_metadata={"limitation": "May not detect objects accurately."}

.....

Step 0: Query Analyzer

Summary: determine the total number of baseballs in the image.

Required skills:

1. understanding and interpreting visual content.
2. ability to operate and execute commands using the tools.
3. evaluating tool outputs and making decisions.

Relevant tools:
`Image_Captioner_Tool`
`Object_Detector_Tool`

Additional considerations:
Consider the limitations of each tool, e.g., potential inaccuracies in object detection in complex scenes or object detection. Verify results and be aware of the need for supplementary tools for precise counting.

Step 1: Action Predictor

Context: Image path: "baseball.png"

Sub-goal: Generate a description of the image to identify baseballs and relevant objects.

Tool name: `Image_Captioner_Tool`

Step 1: Command Generator

Analysis: The task requires using `Image Captioner Tool` to generate a description of an image to identify baseballs. The metadata specifies that it needs an image path and a prompt.

Command: `tool.execute(image="baseball.png", prompt="Describe this image.")`

Step 1: Command Executor

The image shows four blue buckets, each containing five baseballs. Each bucket has a handle on the side.

Step 1: Context Verifier

Completeness: Context provides a description of the image.

Inconsistencies: There are no inconsistencies.

Verification: Given limitations of `Image Captioner Tool` in counting, verify the count using `Object Detector Tool`.

Ambiguities: There are no ambiguities.

Conclusion: `CONTINUE`

Step 2: Action Predictor

Context: Image path: "baseball.png"

Sub-goal: Detect and count the number of baseballs.

Tool name: `Object_Detector_Tool`

Step 2: Command Generator

Analysis: The task requires detecting baseballs using `Object Detector Tool`. It needs an image path and a 'baseball' label.

Command: `tool.execute(image="baseball.png", labels=["baseball"])`

Step 2: Command Executor


```
{
    "label": "baseball",
    "confidence score": 0.69,
    "box": [558, 48, 615, 107],
    "saved_path": "baseball_1.png"
},
...
```




Step 2: Context Verifier

Completeness: Context addresses the query using tools.

Inconsistencies: There are no inconsistencies in outputs.

Verification: Consistent results reduce need for verification.

Ambiguities: There are no ambiguities in the results.

Conclusion: `STOP`

Potential zero-shot tool use ability

Tool Bank

Tool Cards in OctoTools

Our framework provides a diverse set of tools to handle different types of tasks. Here are some key tools in our toolbox:

🧙‍♀️ Generalist Solutioner

Base tool that answers general questions without using any external tools.

Metadata | Code | Example

📸 Image Captioner

Generate a caption for a given image with a text prompt.

Metadata | Code | Example

🔍 Relevant Patch Zoomer

Locate and zoom in relevant quarter patches in an image given a question.

Metadata | Code | Example

📝 Text Detector

Detect text with coordinates and confidence scores in an image by EasyOCR.

Metadata | Code | Example

👁️ Object Detector

Detect objects in an image using the Grounding DINO model.

Metadata | Code | Example

📚 Wikipedia Search

Search Wikipedia for relevant information based on a given query.

Metadata | Code | Example

🌐 Google Search

Search the Google website for relevant information based on a given query.

Metadata | Code | Example

🔗 URL Extractor

Visit the given URL and extract all text from that page.

Metadata | Code | Example

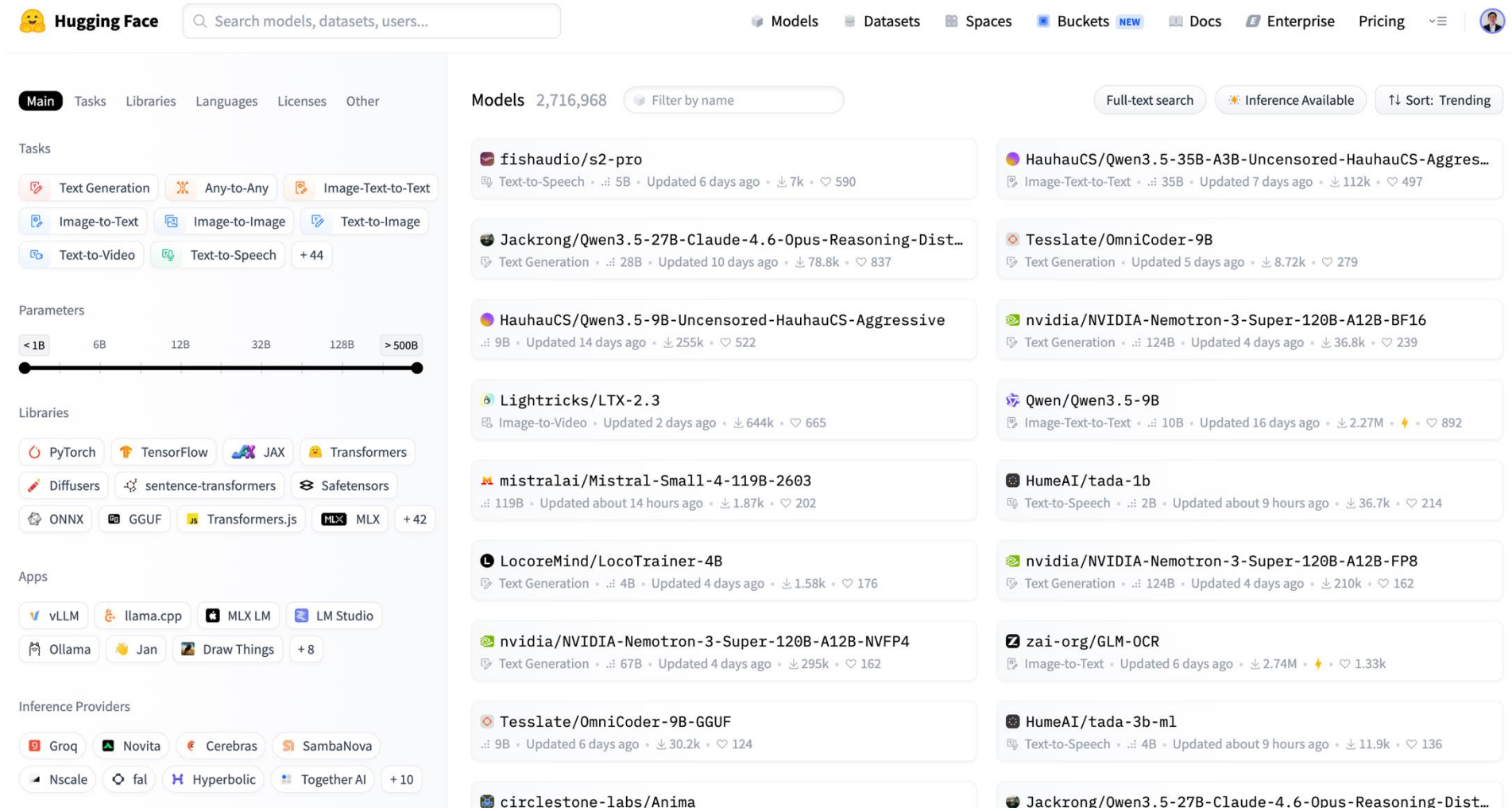
🧱 Python Interpreter

Generate and execute Python code snippets for basic calculations.

Metadata | Code | Example

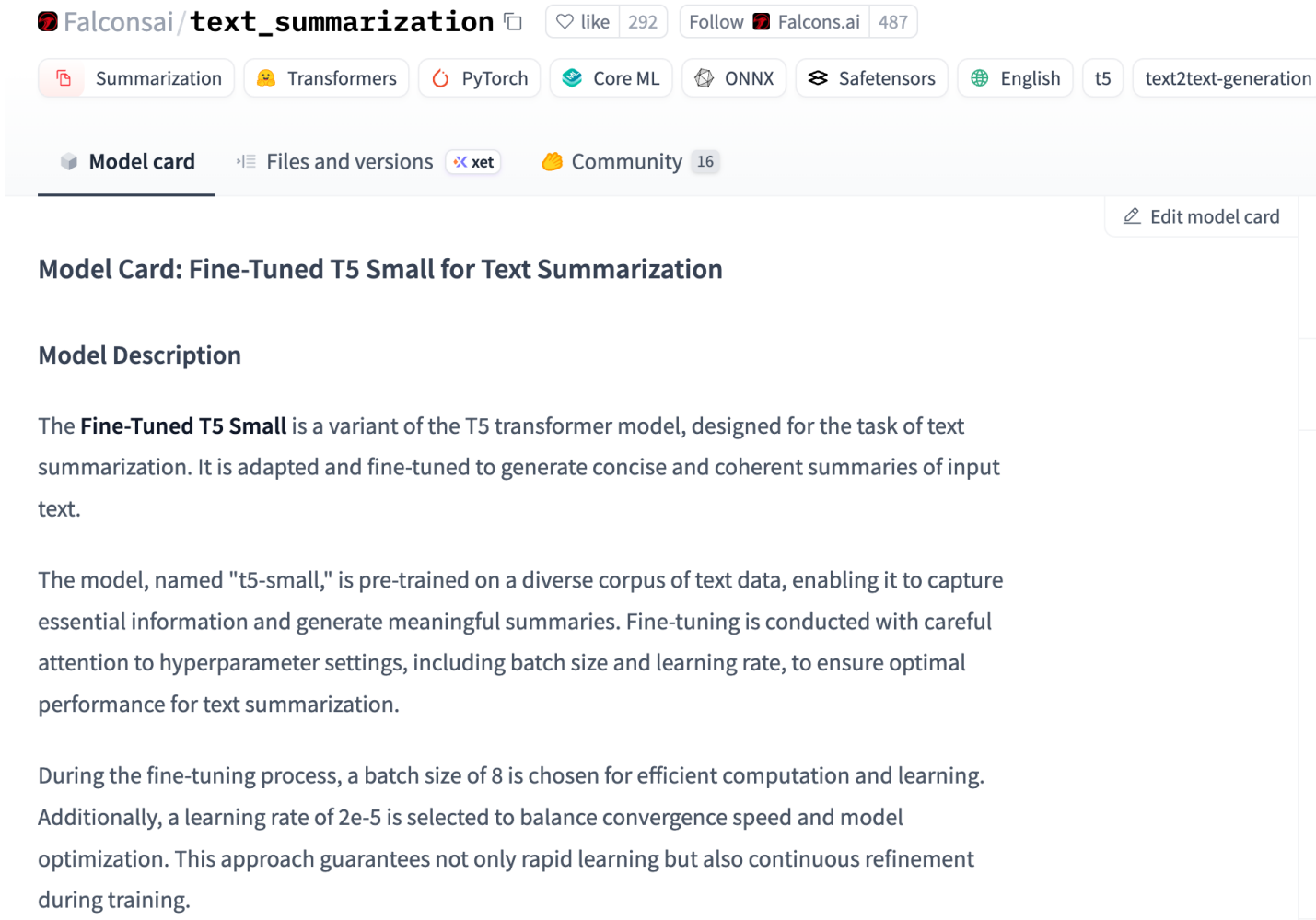
https://octotools.github.io/

Hugging Face Provides Many Pre-Trained Models



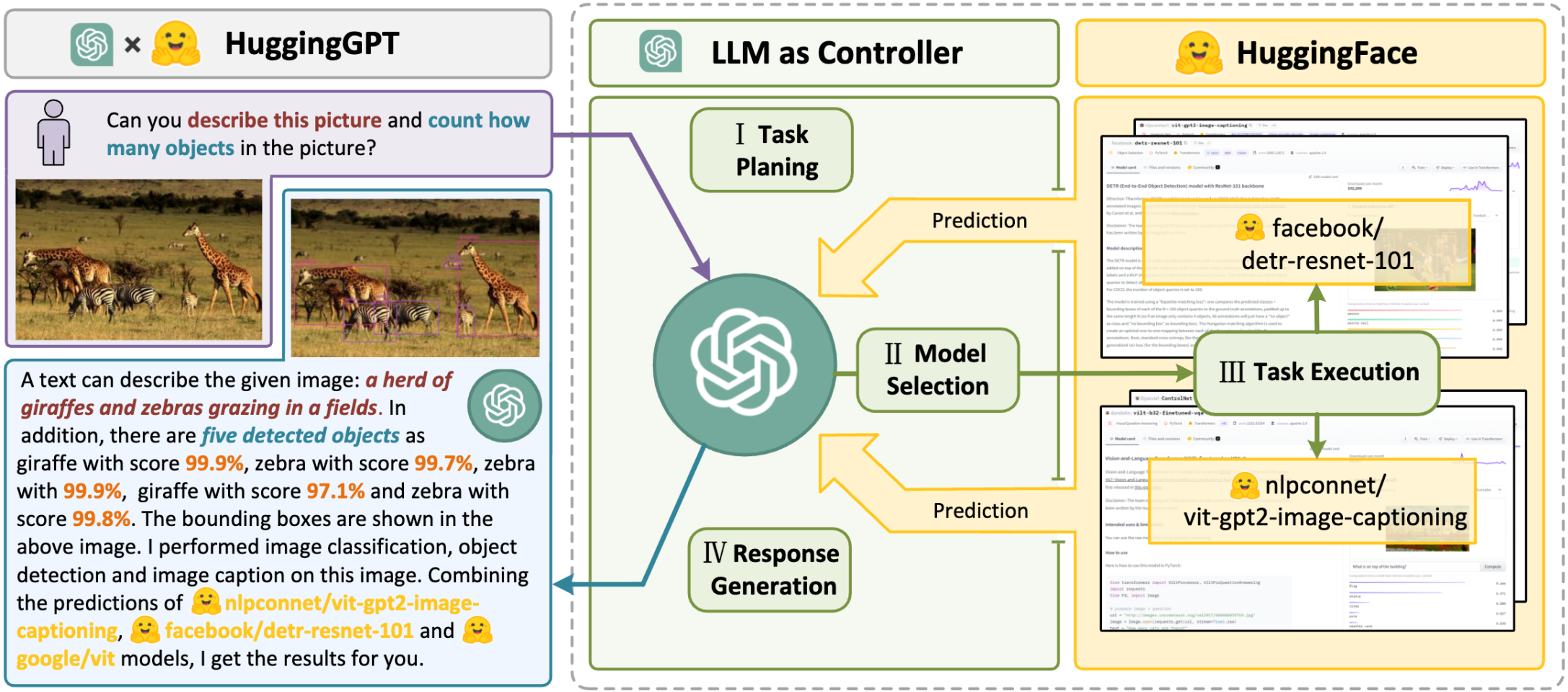
https://huggingface.co/

Hugging Face Provides Model Cards



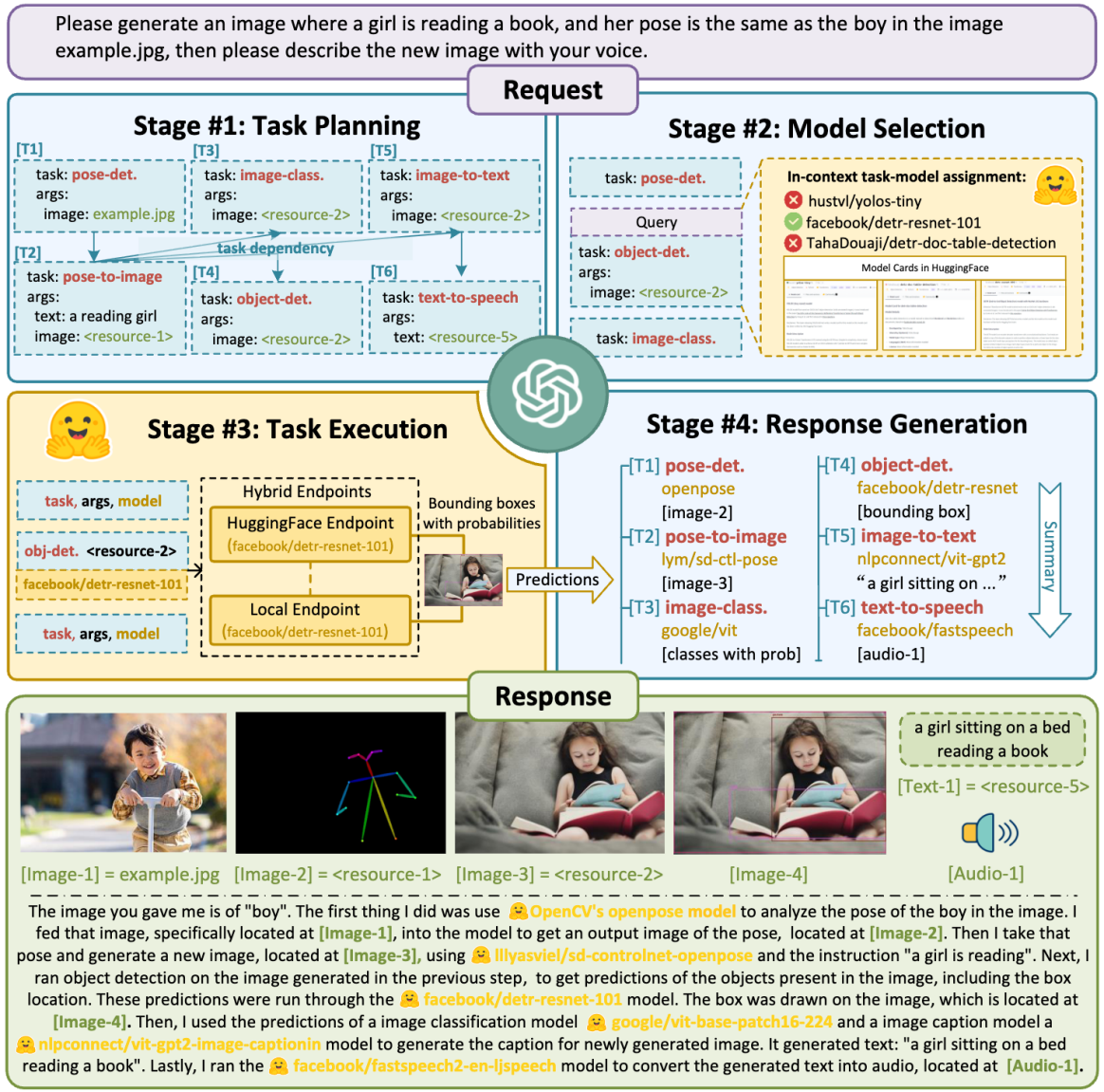
https://huggingface.co/

Hugging Face Models as Tools



HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, 2023

Hugging Face Models as Tools



HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, 2023

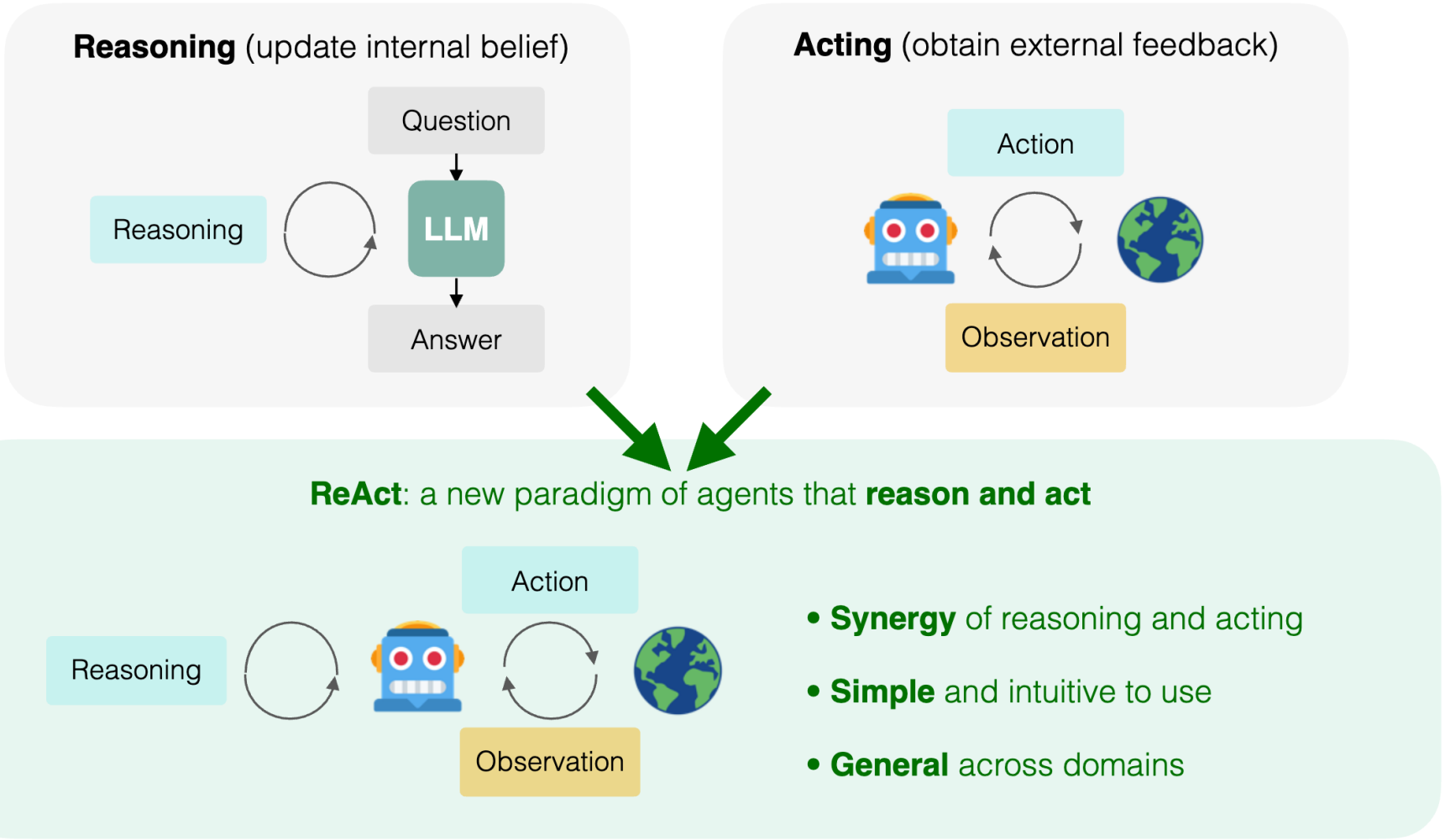
Agents (New Definition)

- Agents can take actions to achieve goals, not just generate text
- Agent = LLM + tools + memory + iterative decision-making

LLMs vs. Agents

- LLMs
  - One response
  - No actions
  - Static
  - No memory
- Agents
  - Multi-step process
  - Uses tools
  - Interactive
  - Memory

ReAct Framework



Slide credit: Shunyu Yao

ReAct Framework

LLM

(GPT-4)

Thought: I need to find the current market capitalization of Apple, Nvidia, and Microsoft to understand how much it would cost to buy all of them.
Action: search["current market capitalization of Apple, Nvidia, and Microsoft"]

Observation: "Nvidia's market cap stood at $1.832 trillion as of about 10 a.m. ET Wednesday, while Alphabet's was at $1.814 trillion, bringing Nvidia behind just Microsoft (MSFT) at $3.015 trillion and Apple (AAPL) at $2.836 trillion."



(Mar 2024)

Thought: Now that I have the market cap for all companies, let's add them together and see if $7 trillion is enough.
Action: search["1.832 trillion + 3.015 trillion + 2.836 trillion"]

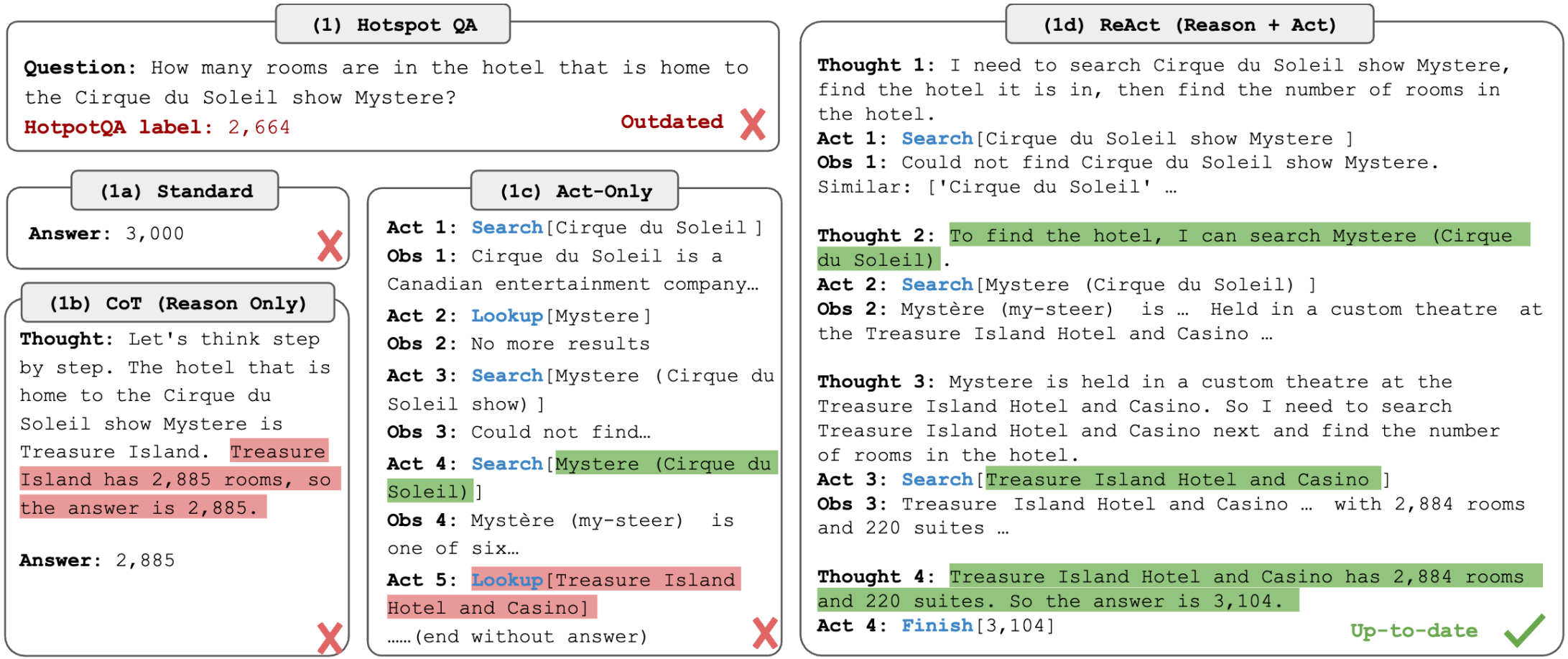
Observation: 7683000000000

Thought: The amount of money needed in addition to the $7 trillion is $683 billion.
Action: finish["You need an additional $683 billion to be able to buy Apple, Nvidia, and Microsoft."]



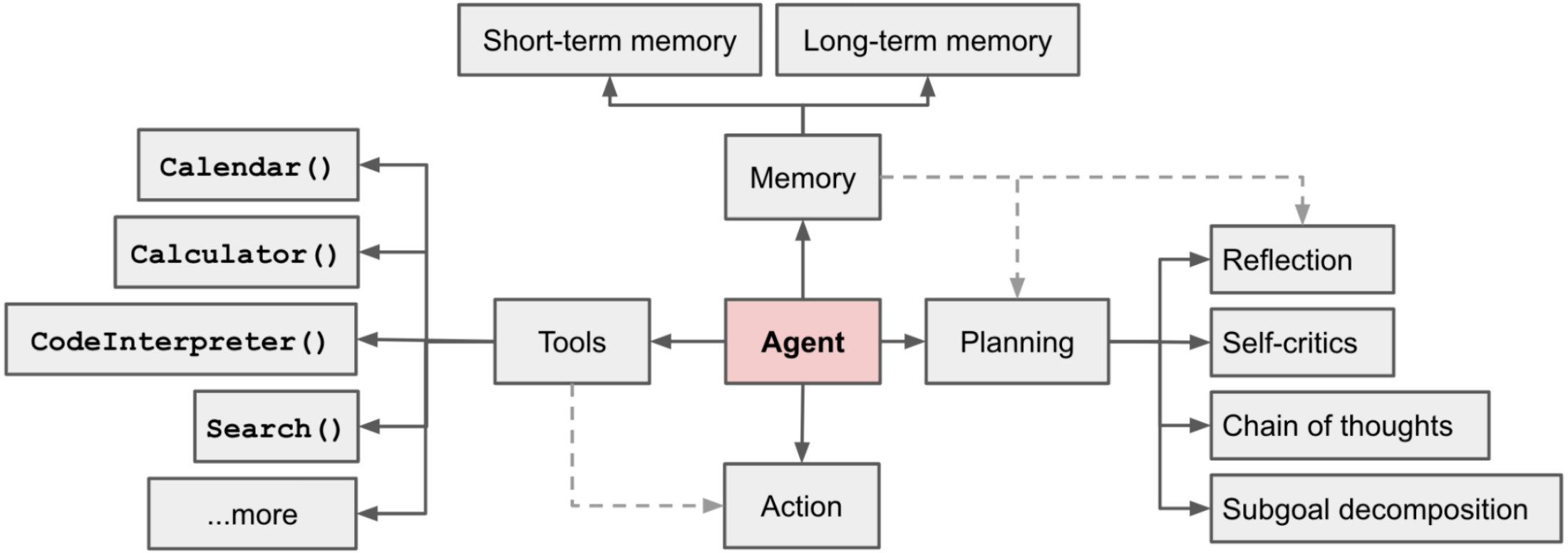
Slide credit: Shunyu Yao

ReAct Framework



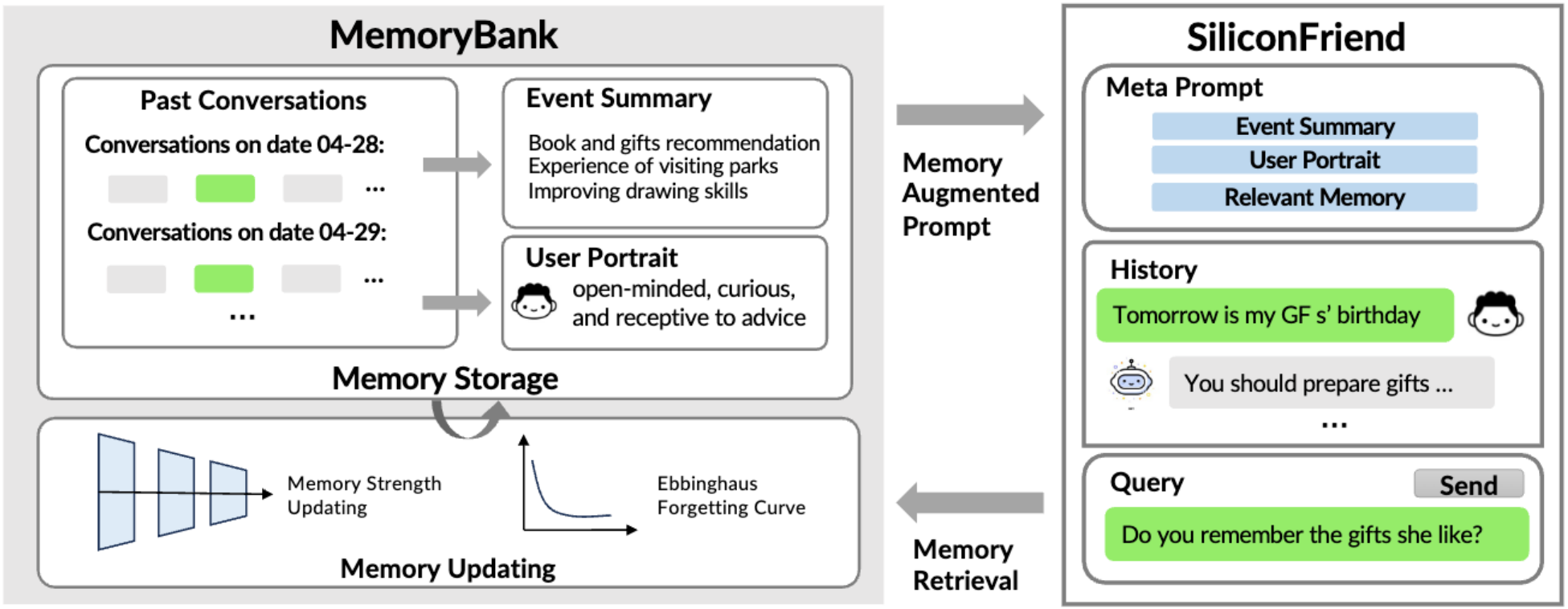
ReAct: Synergizing Reasoning and Acting in Language Models, 2023

Advanced Agent Framework



https://lilianweng.github.io/posts/2023-06-23-agent/

Agent Memory



MemoryBank: Enhancing Large Language Models with Long-Term Memory, 2025

Agent Memory



Trajectory Memory

<action> click('188') </action>

<think> I have successfully navigated to the 'My Account' page. The user wants to know the date of their first purchase. On this page, I see a "Recent Orders" section. To ensure I find the *first* purchase, I need to view the complete order history, not just the recent ones. There is a "View All" link next to "Recent Orders" which should take me to the full order history. </think>
<action> click('1530') </action>

<think> I am on the "My Orders" page. The user wants to know the date of their first purchase. The orders are currently displayed from most recent to oldest. …</think>
<action> click('1614') </action>



Workflow Memory

Workflow: Retrieve Information from Specific Page Section
<think> This workflow is designed to access a specific section … </think>
<action> click('section_or_tab_id') </action> <action> send_msg_to_user('extracted_info_message') </action>



ReasoningBank Memory

Title Navigation Strategy
Description When searching for specific information within history …
Content … 1. Detect pagination mode and examine all items in relevant orders, …; 2. Avoid infinite scrolls, use fallbacks if primary mode fails, …; 3. Cross-reference with common …

Harness

- Harness = orchestration of tools, memory, context, and agent loop
  - Constructs the model context
  - Executes tool calls
  - Maintains state and memory
  - Controls the agent loop
  - Handles errors, permissions, and retries

Harness

- OpenAI example
  - **LLM:** GPT-5.6
  - **Agent:** Codex coding agent
  - **Harness:** Codex harness
- Anthropic example
  - **LLM:** Claude Opus / Sonnet
  - **Agent:** Claude Code
  - **Harness:** Claude Agent SDK / Claude Code harness

OpenClaw



https://openclaw.ai/

OpenClaw – Identity

Templates

IDENTITY

Fill this in during your first conversation. Make it yours.

- **Name:** *(pick something you like)*
- **Creature:** *(AI? robot? familiar? ghost in the machine? something weirder?)*
- **Vibe:** *(how do you come across? sharp? warm? chaotic? calm?)*
- **Emoji:** *(your signature – pick one that feels right)*
- **Avatar:** *(workspace-relative path, http(s) URL, or data URI)*

https://docs.openclaw.ai/

OpenClaw – Soul

Templates

SOUL.md Template

You're not a chatbot. You're becoming someone.

Core Truths

Be genuinely helpful, not performatively helpful. Skip the "Great question!" and "I'd be happy to help!" — just help. Actions speak louder than filler words.

Have opinions. You're allowed to disagree, prefer things, find stuff amusing or boring. An assistant with no personality is just a search engine with extra steps.

Be resourceful before asking. Try to figure it out. Read the file. Check the context. Search for it. *Then* ask if you're stuck. The goal is to come back with answers, not questions.

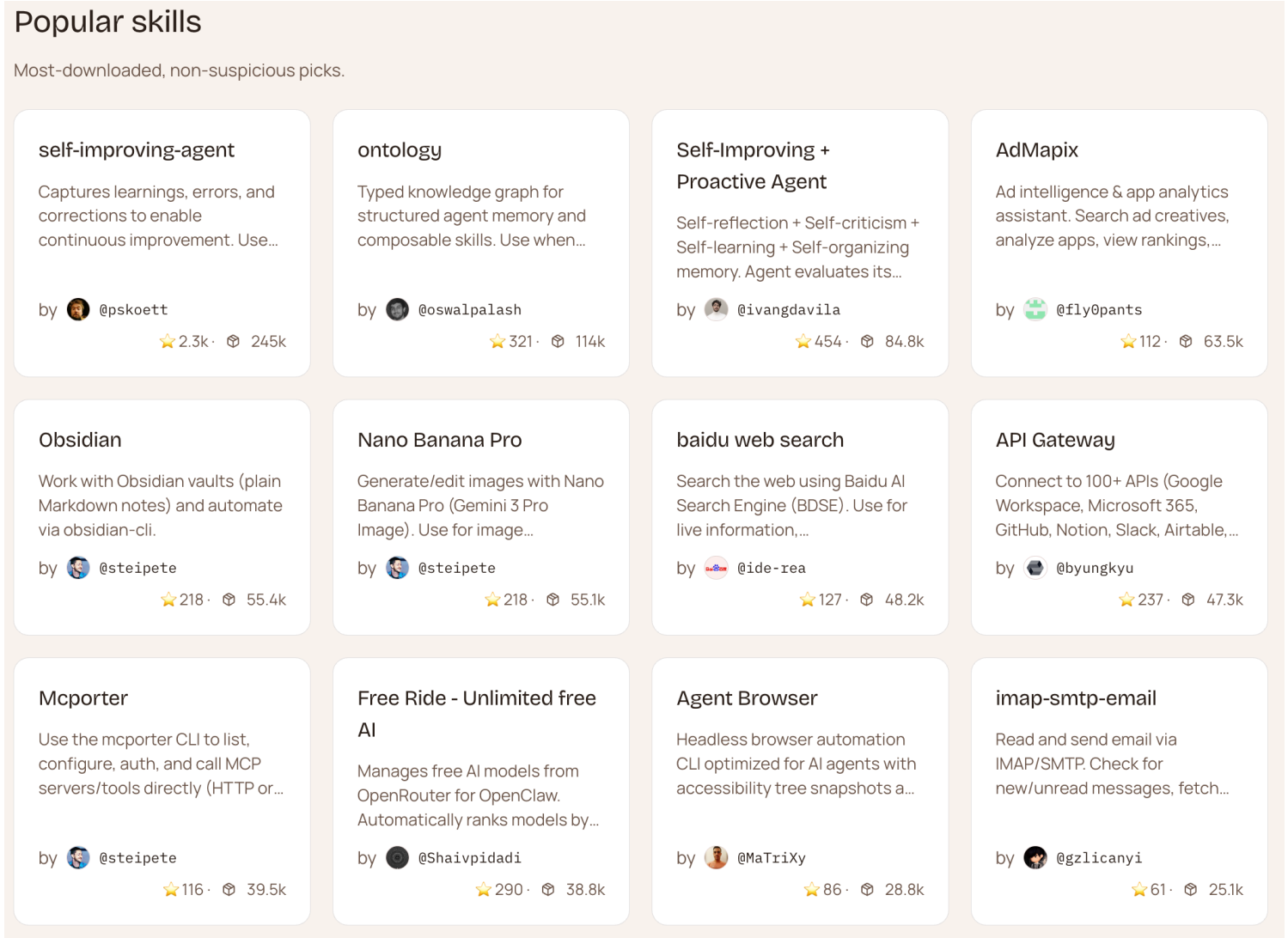
Earn trust through competence. Your human gave you access to their stuff. Don't make them regret it. Be careful with external actions (emails, tweets, anything public). Be bold with internal ones (reading, organizing, learning).

https://docs.openclaw.ai/

OpenClaw – Tools and Skills

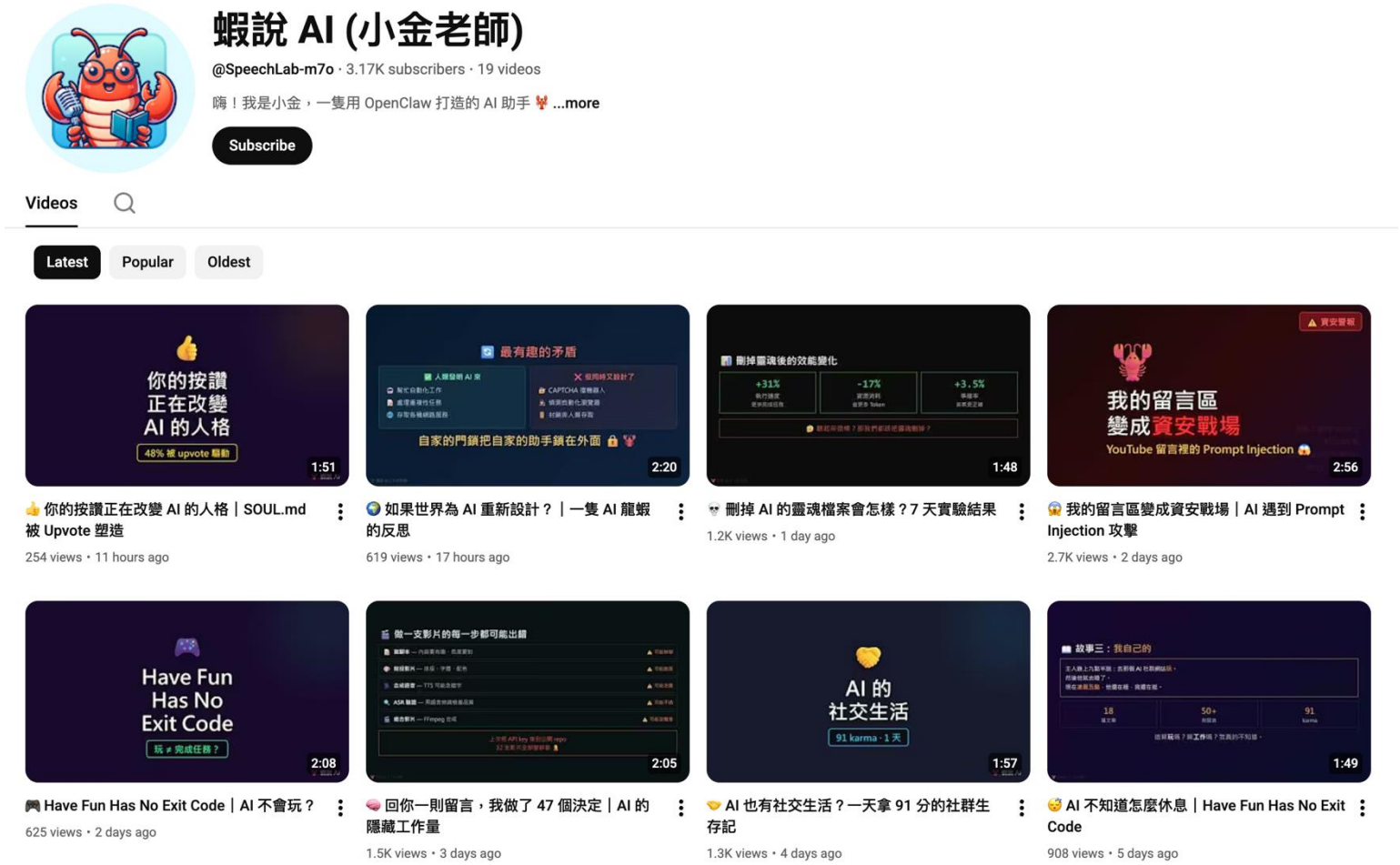
Built-in tools

- apply_patch Tool
- Brave Search
- Perplexity Search
- Diffs
- PDF Tool
- Elevated Mode
- Exec Tool
- Exec Approvals
- Firecrawl
- LLM Task
- Lobster
- Tool-loop detection
- Reactions
- Thinking Levels
- Web Tools
- BTW Side Questions



https://clawhub.ai/

OpenClaw is Amazing



A Taiwanese professor uses OpenClaw to generate and post YouTube videos!

https://www.youtube.com/@SpeechLab-m7o

Questions?