

Agentic AI Overview

CSE 240D

Spring 2026

Hanyang Xu



What is an AI Agent?

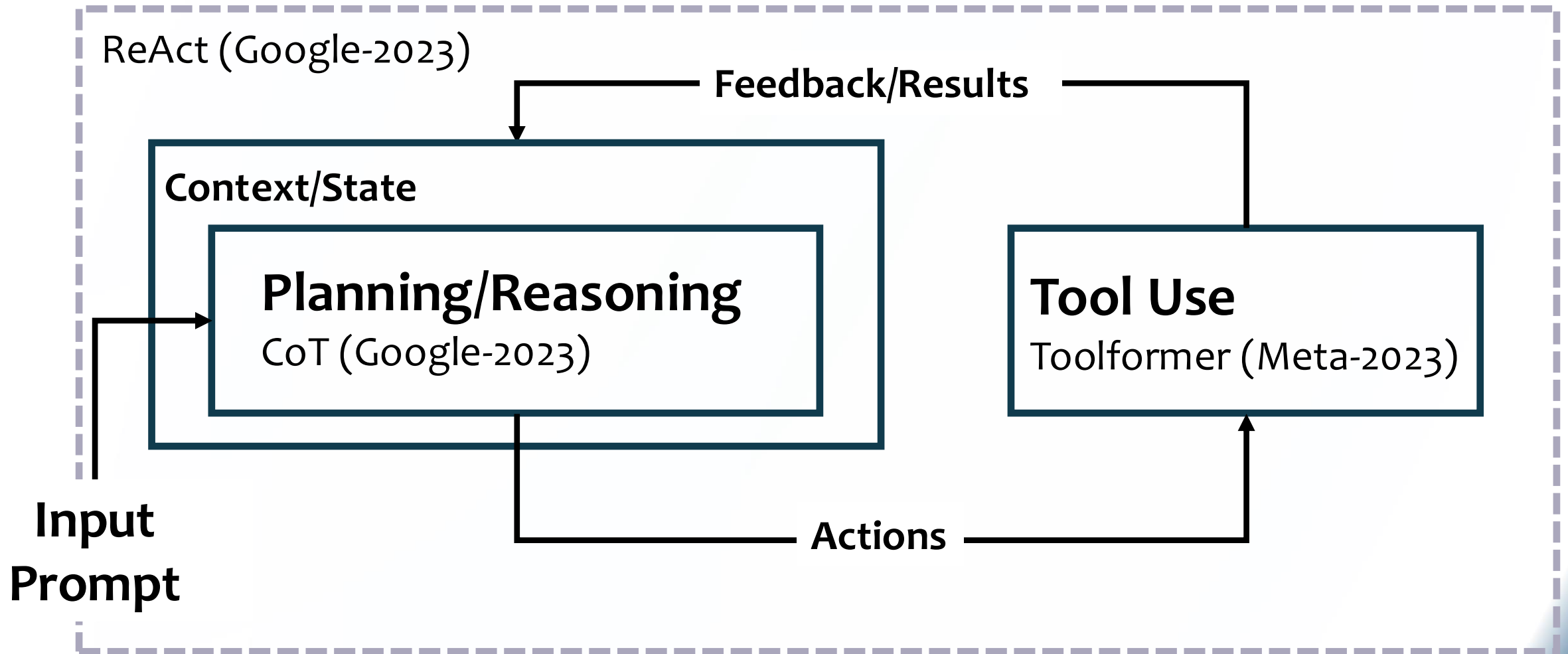
What is an AI Agent?

A Weak Notion of Agency

Perhaps the most general way in which the term agent is used is to denote a hardware or (more usually) software-based computer system that enjoys the following properties:

- **autonomy:** agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state (Castelfranchi, 1995);
- **social ability:** agents interact with other agents (and possibly humans) via some kind of *agent-communication language* (Genesereth and Ketchpel, 1994);
- **reactivity:** agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the INTERNET, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it;
- **pro-activeness:** agents do not simply act in response to their environment, they are able to exhibit goal-directed behaviour by *taking the initiative*.

LLM Agent



Tool Using Through API Calls

Toolformer Tools/APIs

What other name
is Pittsburgh
known by?



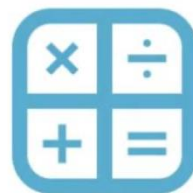
The Steel City

War memorial
Flodden



[...] was created in
memory of the
Battle of Flodden.

$3435 \times$
 $235 / 9$



89691.67

Ø



Thursday,
March 10,
2019

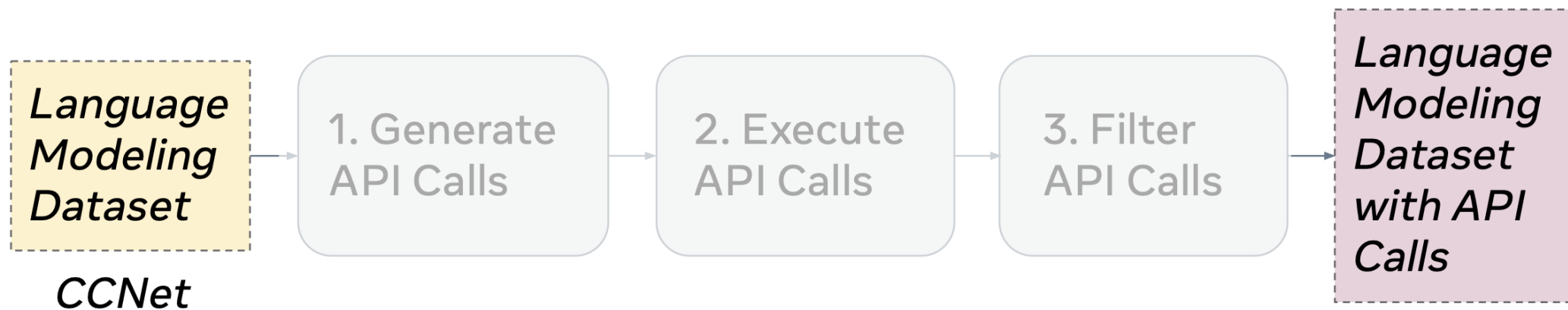
Os Melhores
Escolas em
Jersey



The Best
Schools in
Jersey

Credit to Jane Dwivedi-Yu, Meta

Tool Using Through API Calls



Credit to Jane Dwivedi-Yu, Meta

Self-supervised API Call generation

Prompting the Model to Generate API Calls

Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:

Input: Joe Biden was born in Scranton, Pennsylvania.

Output: Joe Biden was born in [QA("Where was Joe Biden born?")]
Scranton, [QA("In which state is Scranton?")] Pennsylvania.

Input: $\${input}$

Output:

Usefulness by Perplexity

Filter the API Calls using Model-based Perplexity

$$L_{\bullet}(PREFIX) = -\log p(\text{the Steel City.} \mid PREFIX)$$

A. No API Call $L_A(\text{Pittsburgh is known as}) = 2.5$

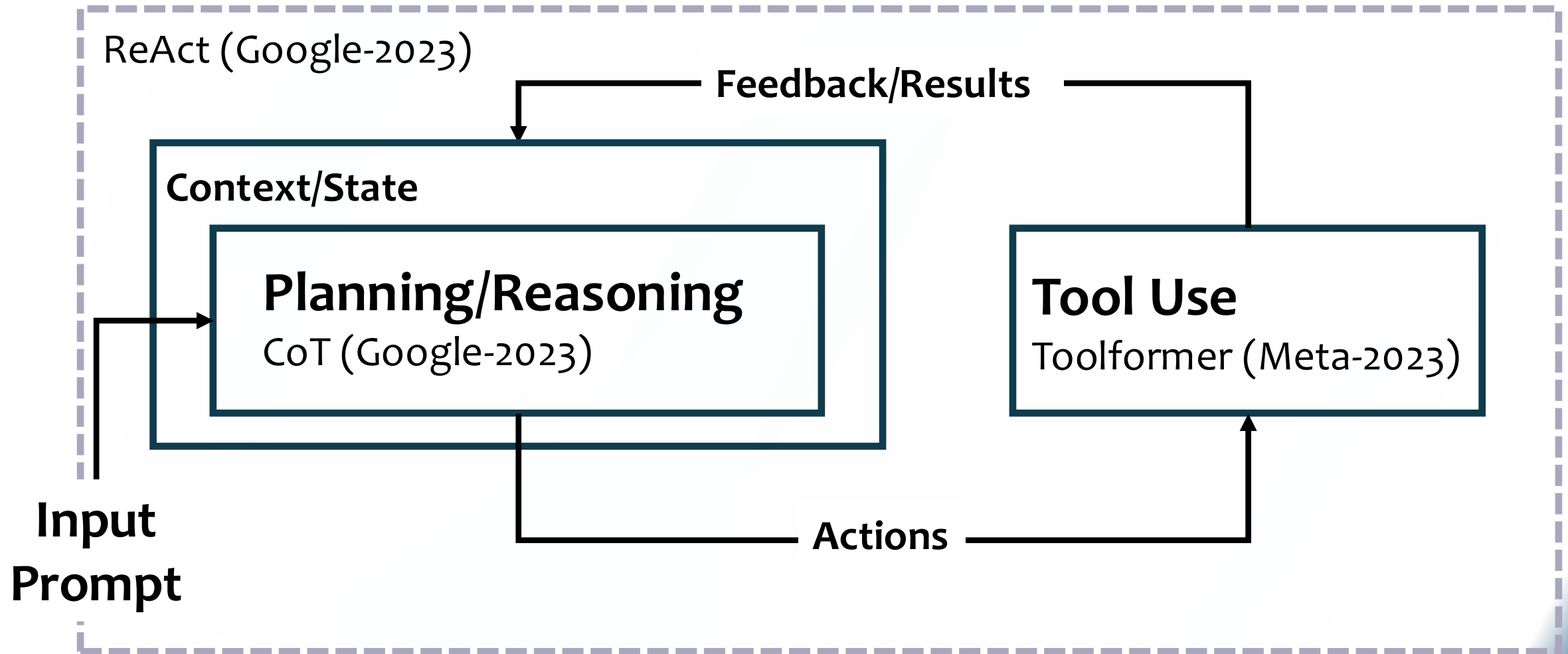
B. Non-executed API Call $L_B(\text{Pittsburgh is known as } [QA(\text{"What other name is Pittsburgh known by?"}) \rightarrow ?]) = 2.1$

C. Executed API Call $L_C(\text{Pittsburgh is known as } [QA(\text{"What other name is Pittsburgh known by?"}) \rightarrow \text{Steel City}]) = 0.8$

$$\text{Usefulness} = \min(L_A, L_B) - L_C = \min(2.5, 2.1) - 0.8 = 1.3$$

Credit to Jane Dwivedi-Yu, Meta

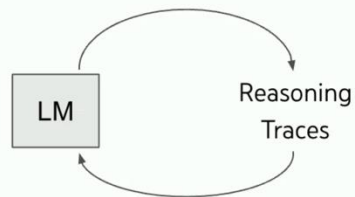
ReAct: Synergizing Reasoning and Acting in Language Models



ReAct: Synergizing Reasoning and Acting in Language Models

Reasoning and Acting

- Language models (LM) get better at reasoning and acting, **separately**



Reason Only

e.g. Chain-of-thought (Wei et al.), Self-consistency (Wang et al.)

Input: 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.

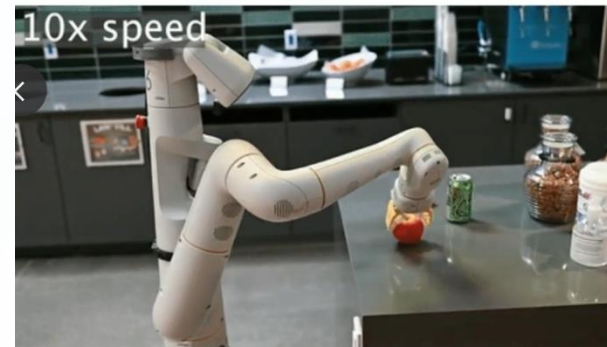
...

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
A:

Model output: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

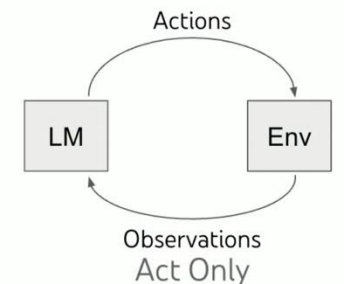
Reasoning and Acting

- Language models (LM) get better at reasoning and acting, **separately**



User input: I just worked out, can you bring me a drink and a snack to recover?

Robot: I would 1. find a water bottle, 2. pick up the water bottle, 3. bring it to you, 4. put down the water bottle, 5. find an apple



e.g. SayCan (Ahn et al.), WebGPT (Nakano et al.)

Credit to Shunyu Yao, Google

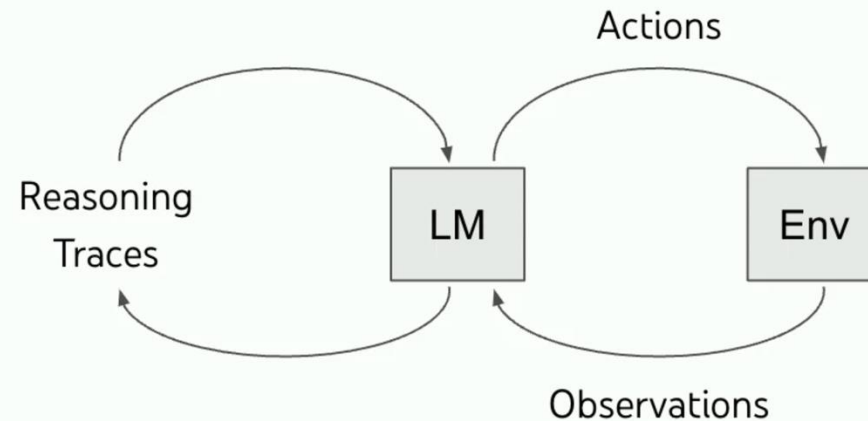
ReAct: Synergizing Reasoning and Acting in Language Models

Reasoning and Acting

Language models (LM) get better at reasoning and acting, **separately**

ReAct: Let LMs generate both

- **Reasoning traces** (to update **internal** model belief)
- **Actions** (to obtain **external** environment feedback)



Credit to Shunyu Yao, Google

ReAct Loop through In-Context Learning

ReAct prompt

Question Which magazine was started first Arthur's Magazine or First for Women?

Thought 1 I need to search Arthur's Magazine and First for Women, and find which was started first.

Action 1 Search[Arthur's Magazine]

Observation 1 Arthur's Magazine (1844-1846) was an American literary periodical published in Philadelphia ...

Thought 2 Arthur's Magazine was started in 1844. I need to search First for Women next.

Action 2 Search[First for Women]

Observation 2 First for Women is a woman's magazine published by Bauer Media Group ... was started in 1989...

Thought 3 First for Women was started in 1989. 1844 (Arthur's Magazine) < 1989 (First for Women), so Arthur's Magazine was started first.

Action 3 Finish[Arthur's Magazine]

Credit to Shunyu Yao, Google

On Going Researches on Agents

Environment: *Physical World, Chip Layout, GUI, Coding Sandbox, etc*

Feedback/Results:

Multi-Agent Communications

Context/State: *How to Manage*

Planning/Reasoning:

*Pretrain from trace
Multi-Agent*

Tool Use:

SQL, VectorDB, git, grep

Actions