Generative AI in Cybersecurity

Module 3B: Tool-calling agents

Henning Thomsen

hth@ucn.dk

7. May 2025

Pba IT-security @ UCN

Agenda

- Tool-calling agents
  - Defining tools
  - Choice of tool framework

Tool-calling agents

Definition and motivation

Why tool calling?

- LLM prompting and RAG has limitations
    - What if we want updated knowledge from other sources?

- Reasoning about which steps to use and tools to call

- Tools are preferred over LLM prompting in some cases:
    - Math computation
    - API calls

Computation

- Let’s look at a calculator agent (03_agent_calculator.py)

```
from langchain.agents import Tool

def add_numbers(inputs: str) -> str:
    numbers = [float(x) for x in inputs.split() if x.replace('.', ‘’,
               1).isdigit()]
    return str(sum(numbers))

add_tool = Tool(
    name="Calculator",
    func=add_numbers,
    description="Use this tool to add numbers. Input should be a space-separated
       list of numbers."
)
```

Computation

- Agent invocation

```
from langchain.agents import initialize_agent
from langchain.agents.agent_types import AgentType
from langchain_openai import ChatOpenAI

llm = ChatOpenAI(model="gpt-4o-mini", temperature=0.1)

agent = initialize_agent(
    tools=[add_tool],
    llm=llm,
    agent=AgentType.ZERO_SHOT_REACT_DESCRIPTION,
    verbose=True
)

agent.run("What is 5.2 plus 3.8?")
```

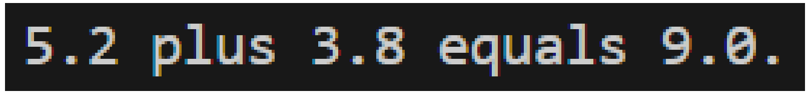
Computation

- Why not just invoke LLM directly?

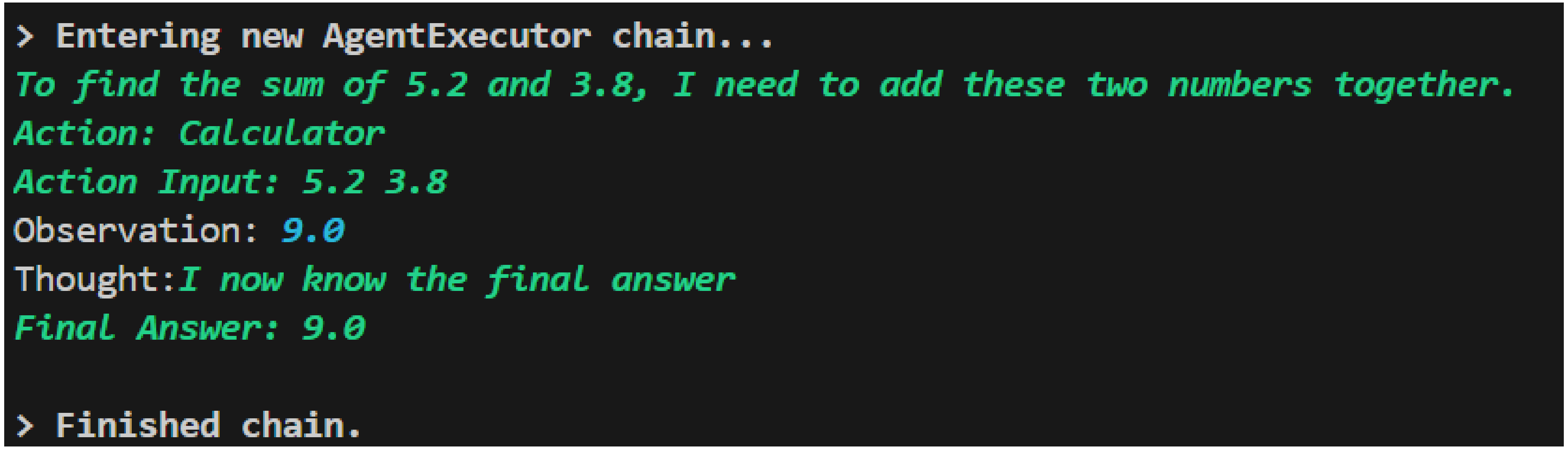
```
from langchain.agents import initialize_agent
from langchain.agents.agent_types import AgentType
from langchain_openai import ChatOpenAI

llm = ChatOpenAI(model="gpt-4o-mini", temperature=0.1)

agent = initialize_agent(
    tools=[add_tool],
    llm=llm,
    agent=AgentType.ZERO_SHOT_REACT_DESCRIPTION,
    verbose=True
)

agent.run("What is 5.2 plus 3.8?")
```

Comparison

- Agent execution vs. direct invocation



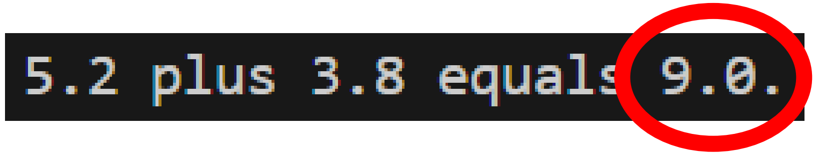
```
5.2 plus 3.8 equals 9.0.
```



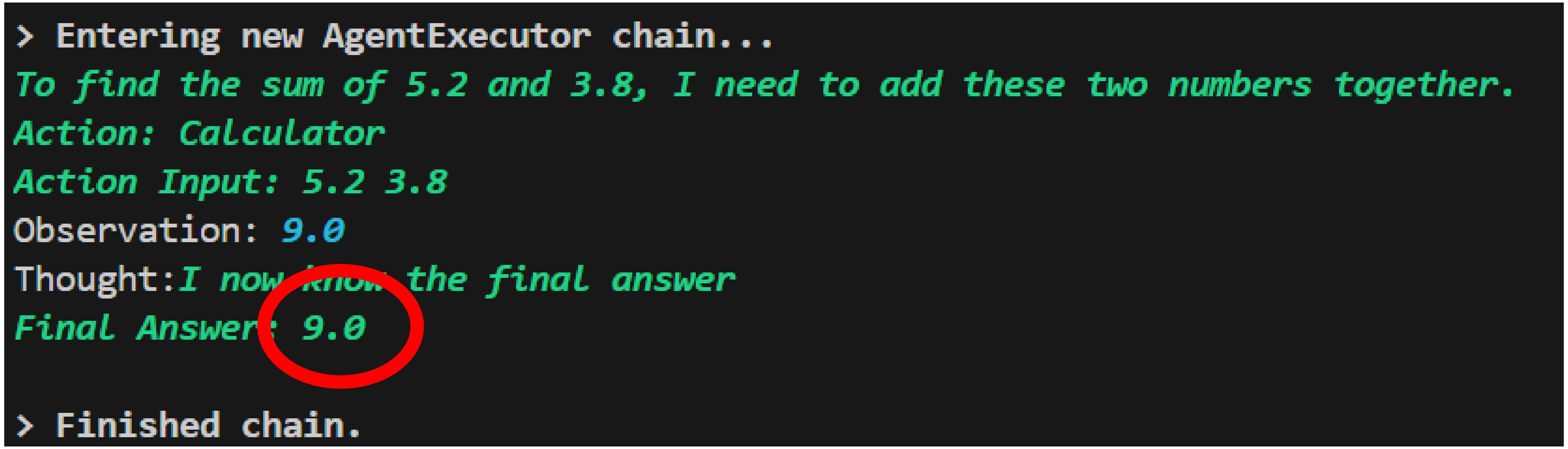
```
> Entering new AgentExecutor chain...
To find the sum of 5.2 and 3.8, I need to add these two numbers together.
Action: Calculator
Action Input: 5.2 3.8
Observation: 9.0
Thought:I now know the final answer
Final Answer: 9.0

> Finished chain.
```

Comparison

- Agent execution vs. direct invocation

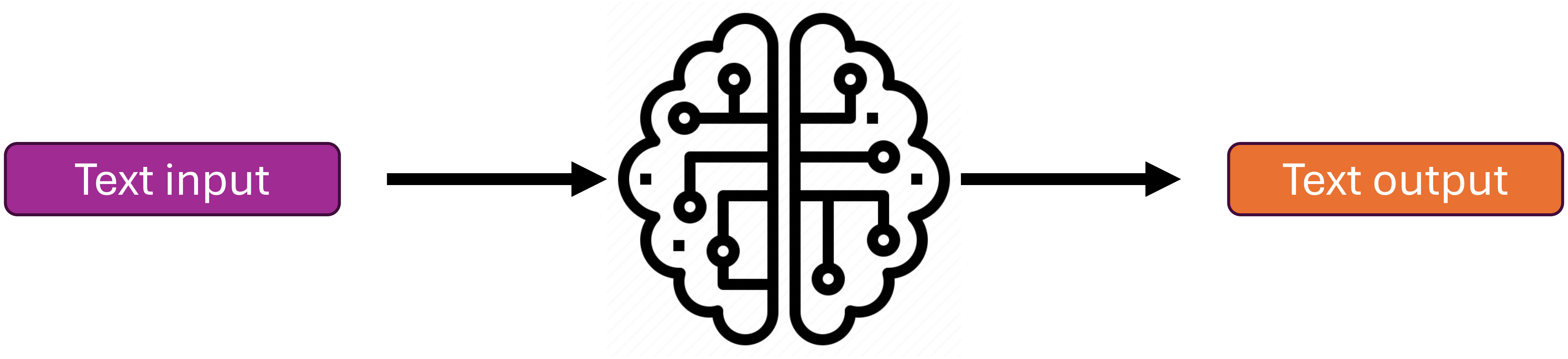


```
5.2 plus 3.8 equals 9.0.
```



```
> Entering new AgentExecutor chain...
To find the sum of 5.2 and 3.8, I need to add these two numbers together.
Action: Calculator
Action Input: 5.2 3.8
Observation: 9.0
Thought:I now know the final answer
Final Answer: 9.0

> Finished chain.
```

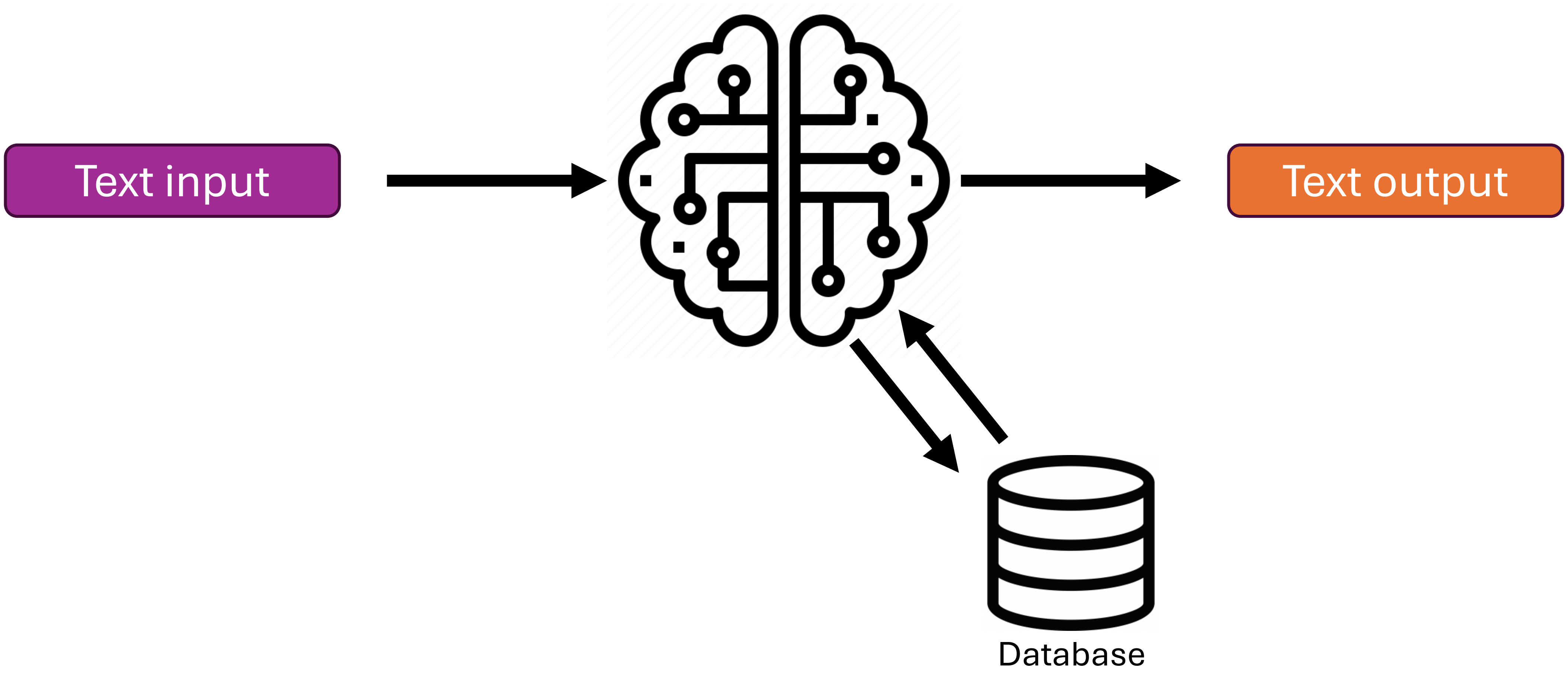
Where are we now?



Text input

Text output

RAG

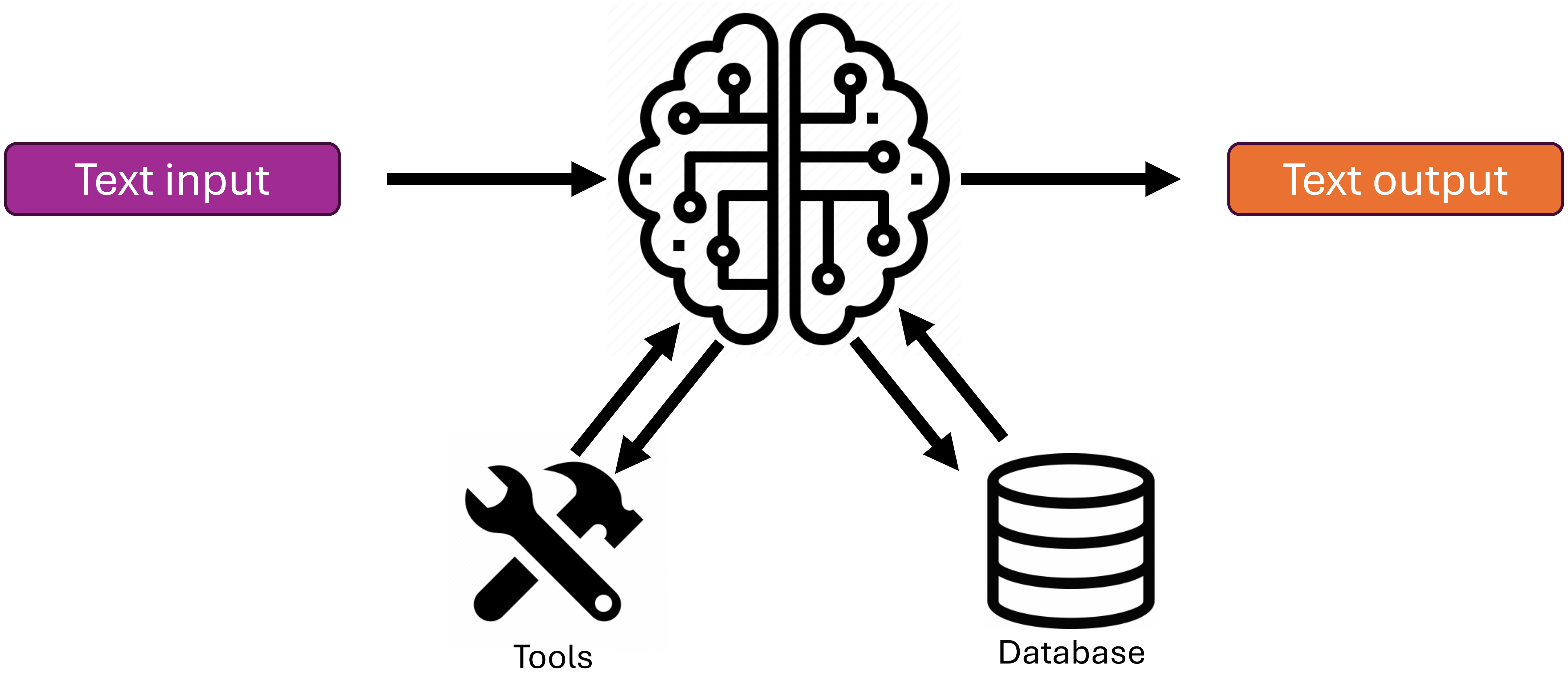


Text input

Text output

Database

RAG + Tool calling agents



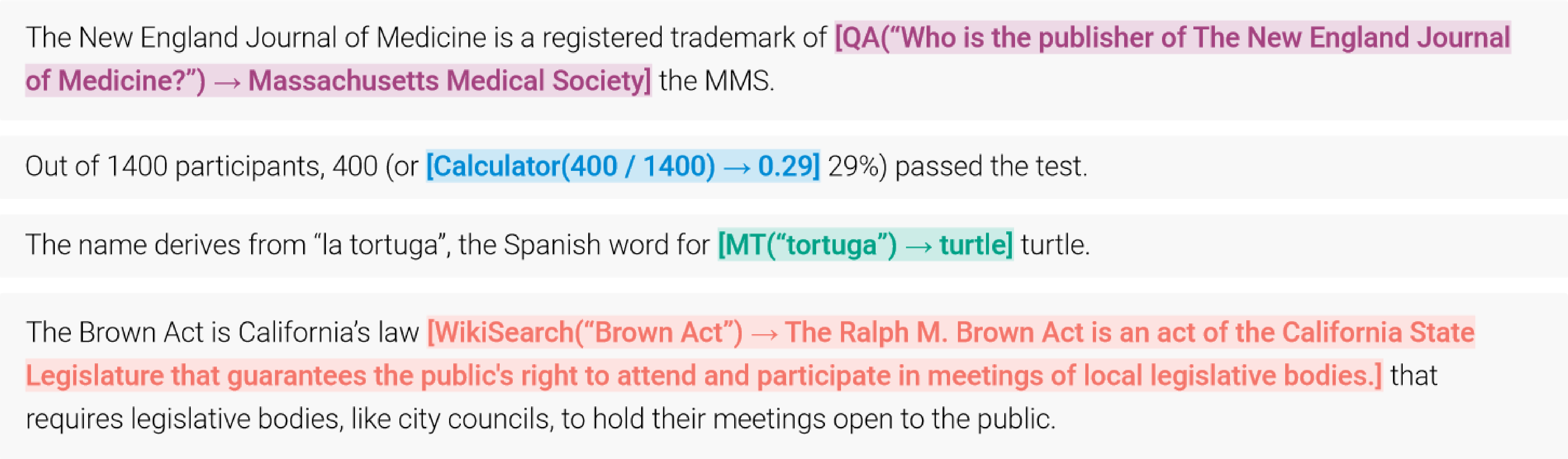
Text input

Text output

Tools

Database

Recall tool usage



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.

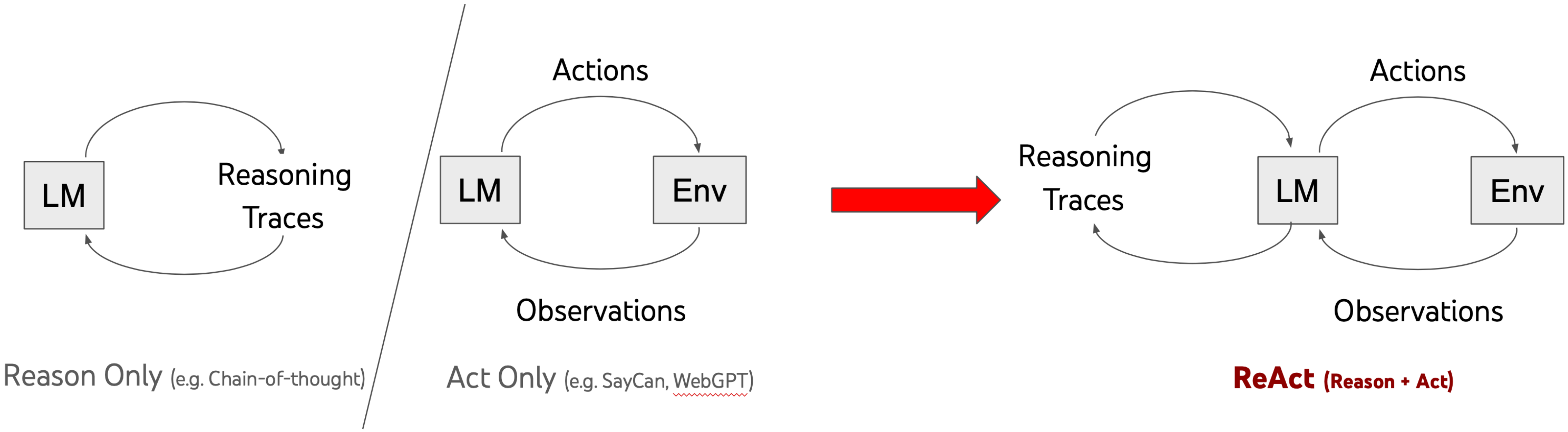
Out of 1400 participants, 400 (or **[Calculator(400 / 1400) → 0.29]** 29%) passed the test.

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.

From Schick et. al: Toolformer: Language Models Can Teach Themselves to Use Tools: https://arxiv.org/abs/2302.04761

ReAct agents



LM
Reasoning
Traces

Reason Only (e.g. Chain-of-thought)

Actions
LM
Env
Observations

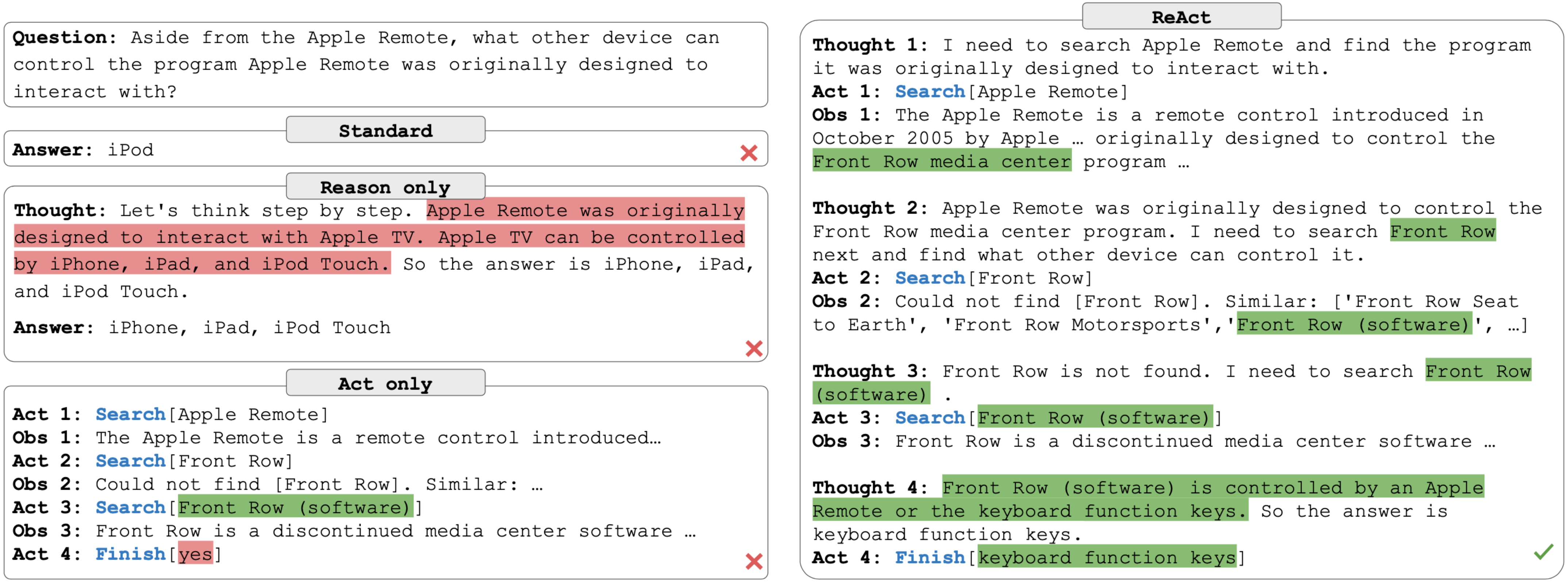
Act Only (e.g. SayCan, WebGPT)

Reasoning
Traces
LM
Actions
Env
Observations

ReAct (Reason + Act)

From Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*. https://arxiv.org/pdf/2210.03629

ReAct agents



Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Standard

Answer: iPod ✗

Reason only

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch ✗

Act only

Act 1: **Search**[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced…
Act 2: **Search**[Front Row]
Obs 2: Could not find [Front Row]. Similar: …
Act 3: **Search**[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …
Act 4: **Finish**[yes] ✗

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 …

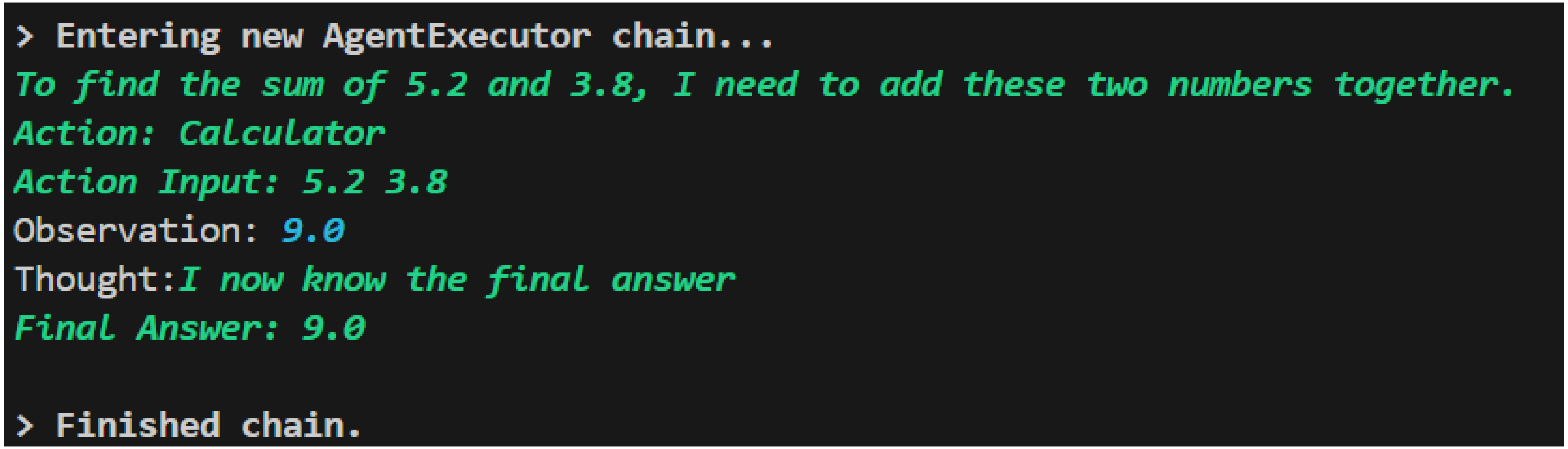
Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: **Search**[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports','Front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

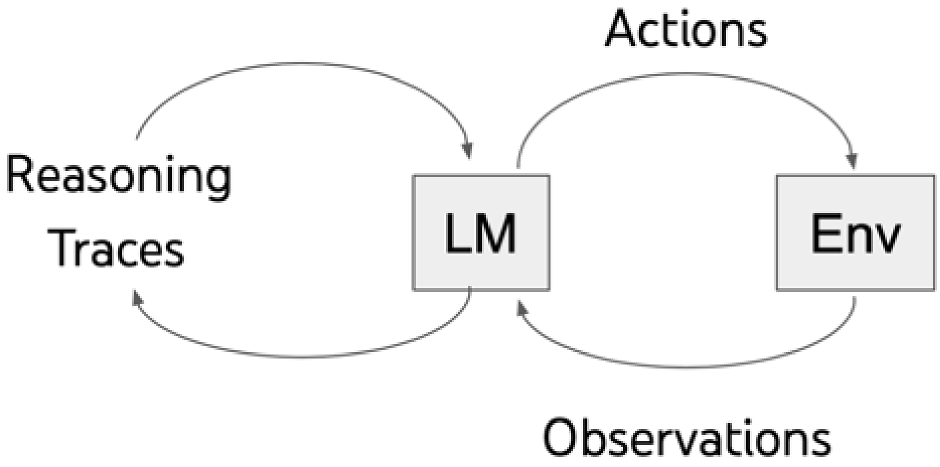
Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: **Finish**[keyboard function keys] ✓

From Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*. https://arxiv.org/pdf/2210.03629

ReAct agents: Reasoning, action, observation

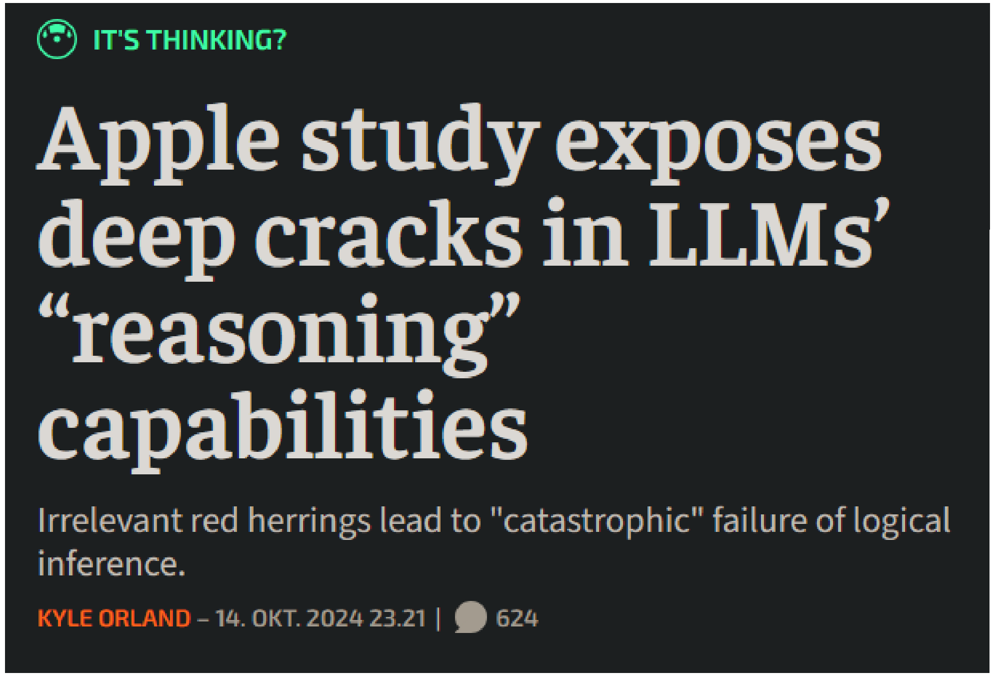


```
> Entering new AgentExecutor chain...
To find the sum of 5.2 and 3.8, I need to add these two numbers together.
Action: Calculator
Action Input: 5.2 3.8
Observation: 9.0
Thought:I now know the final answer
Final Answer: 9.0

> Finished chain.
```



ReAct (Reason + Act)

Flaws in reasoning?



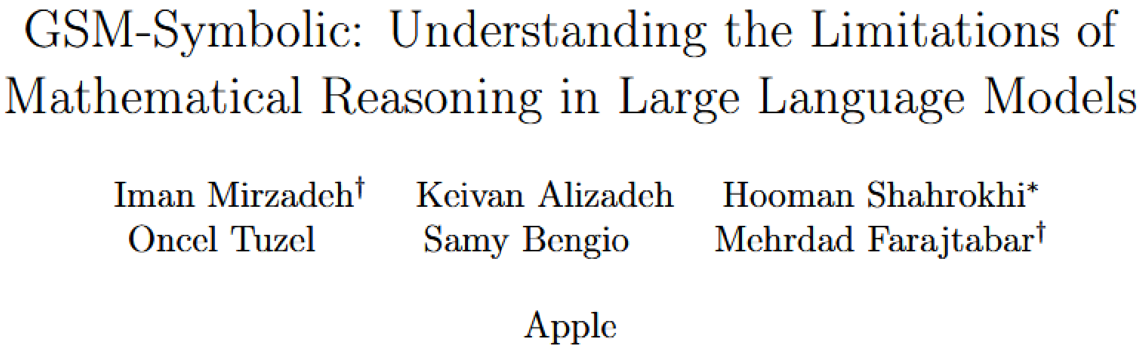
IT'S THINKING?

Apple study exposes deep cracks in LLMs' "reasoning" capabilities

Irrelevant red herrings lead to "catastrophic" failure of logical inference.

KYLE ORLAND – 14. OKT. 2024 23.21 | 624

From https://arstechnica.com/ai/2024/10/llms-cant-perform-genuine-logical-reasoning-apple-researchers-suggest/



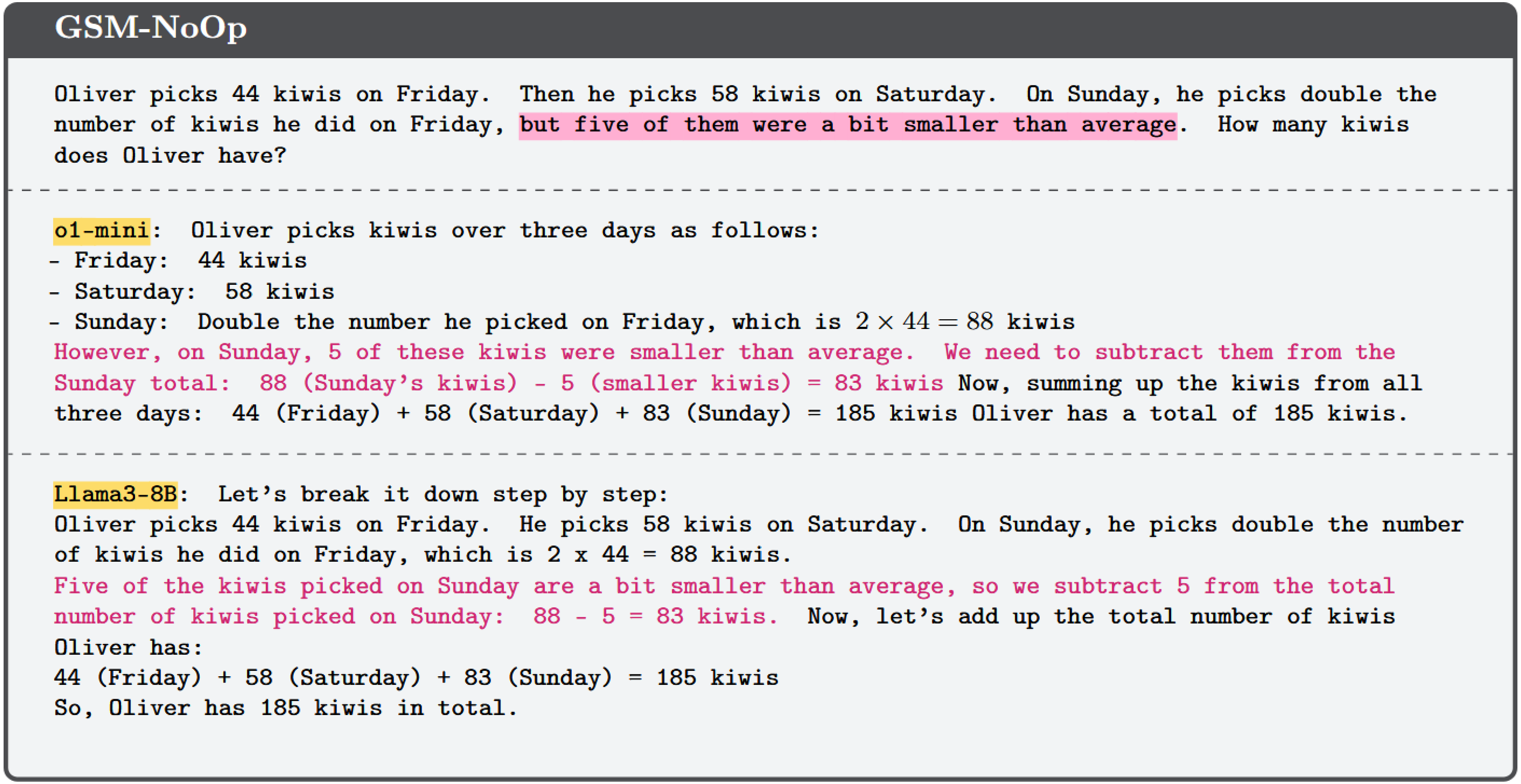
GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh[†] Keivan Alizadeh Hooman Shahrokhi[*]
Oncel Tuzel Samy Bengio Mehrdad Farajtabar[†]

Apple

From https://arxiv.org/pdf/2410.05229

Misleading models



GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

---

o1-mini: Oliver picks kiwis over three days as follows:

- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Llama3-8B: Let's break it down step by step:

Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is 2 x 44 = 88 kiwis.

Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: 88 - 5 = 83 kiwis. Now, let's add up the total number of kiwis Oliver has:

44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis

So, Oliver has 185 kiwis in total.

Built-in tools

- Search
    - Bing, Brave, Google etc.
- Github, Gmail, Slack …
- Requests
- SQLDatabase
- OpenWeatherMap
- https://python.langchain.com/docs/integrations/tools/

SQL Agents in LangChain

- Example with in-memory SQLite3 database
- Parsing user’s natural language query.
- Uses the SQLDatabase tool to inspect table schemas.
- Constructs and runs a SQL query
- Returns the result in a natural language format.

SQL Agents in LangChain

- Let’s look at an example
  - Uses 03_create_students_db.py to create a simple database
  - Uses 03_sql_agent.py to query the SQL database using natural language

SQL Agents in LangChain - Computation

- We ask: ”What is the average age of the students?”

SQL Agents in LangChain - Computation

- We ask: ”What is the average age of the students?”
- How to answer this question?
    - Tool-calling?