Onepoint
TechTalk

Decoding AI series

Session 6

AI agents demystified

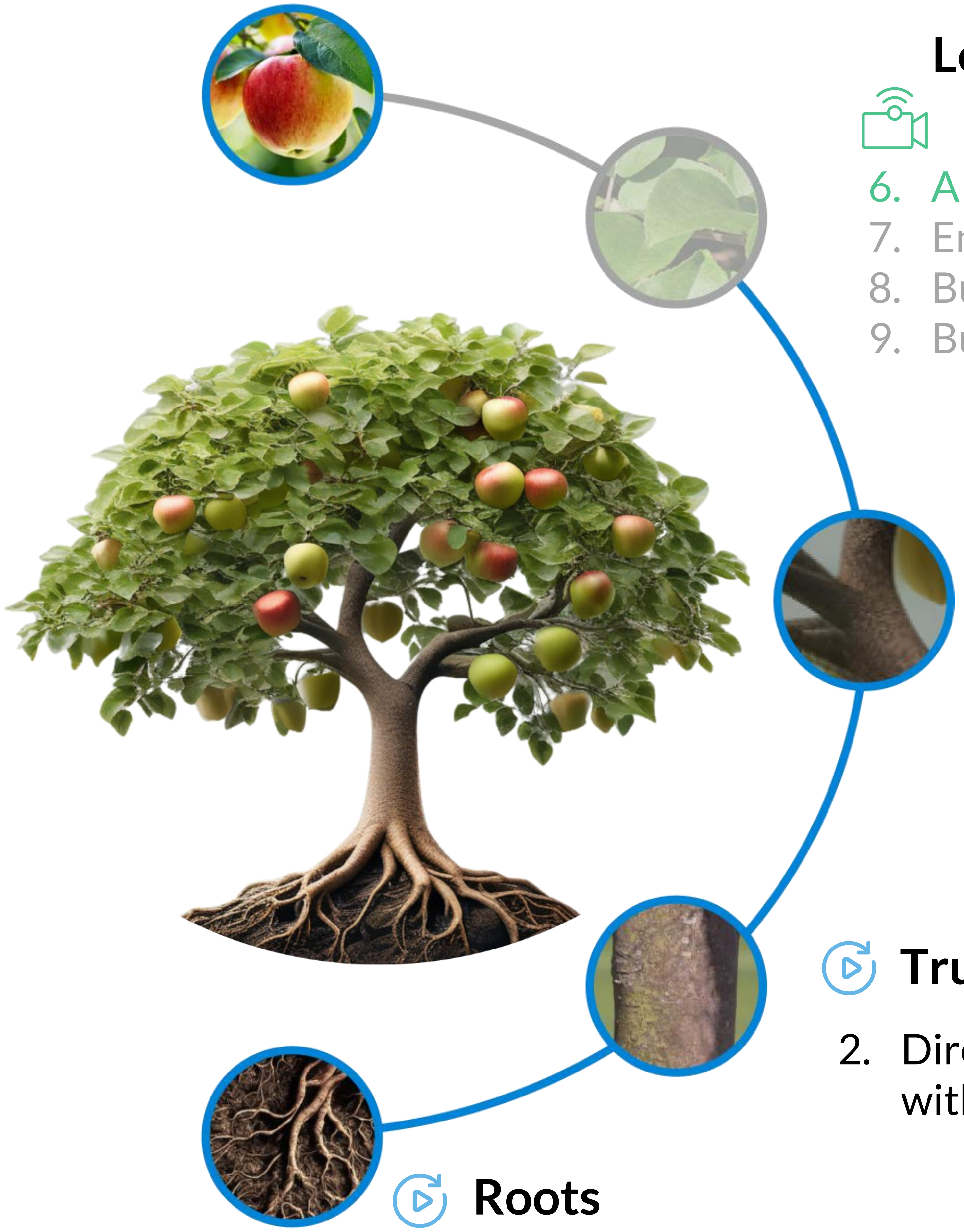
Our awesome speaker

Gil Fernandes
AI Solutions Engineer
Onepoint

Welcome

From Roots to Fruits

- Journey to create value from LLMs



Fruit

3. How to turn the spotlight on Dark Data with Gen AI

Leaves

6. AI agents
7. Ensemble of agents
8. Building complex simulations
9. Building LLMs based systems

Branches

4. Workflows and complex interactions
5. RAG, GraphRAG and beyond

Trunk

2. Direct interactions with LLMs

Roots

1. Fundamentals of LLMs

Agenda

Insights from previous webinar 04

What are agents? 08

Self improving agents 35

What's next? 44

Credits 47

Thank you for joining 48

Insights from previous webinar

Audience poll

Which AI agents are you using?

A. Search agents – agents that perform real time searches
B. Database agents – agents that perform real time queries on databases
C. Personal email agents – agents that perform real time queries on private email
D. Retail agents – agents that perform product searches which help you finding products
E. Price comparison agents – agents that compare prices of retail items
F. Something else
G. No experience with agents yet

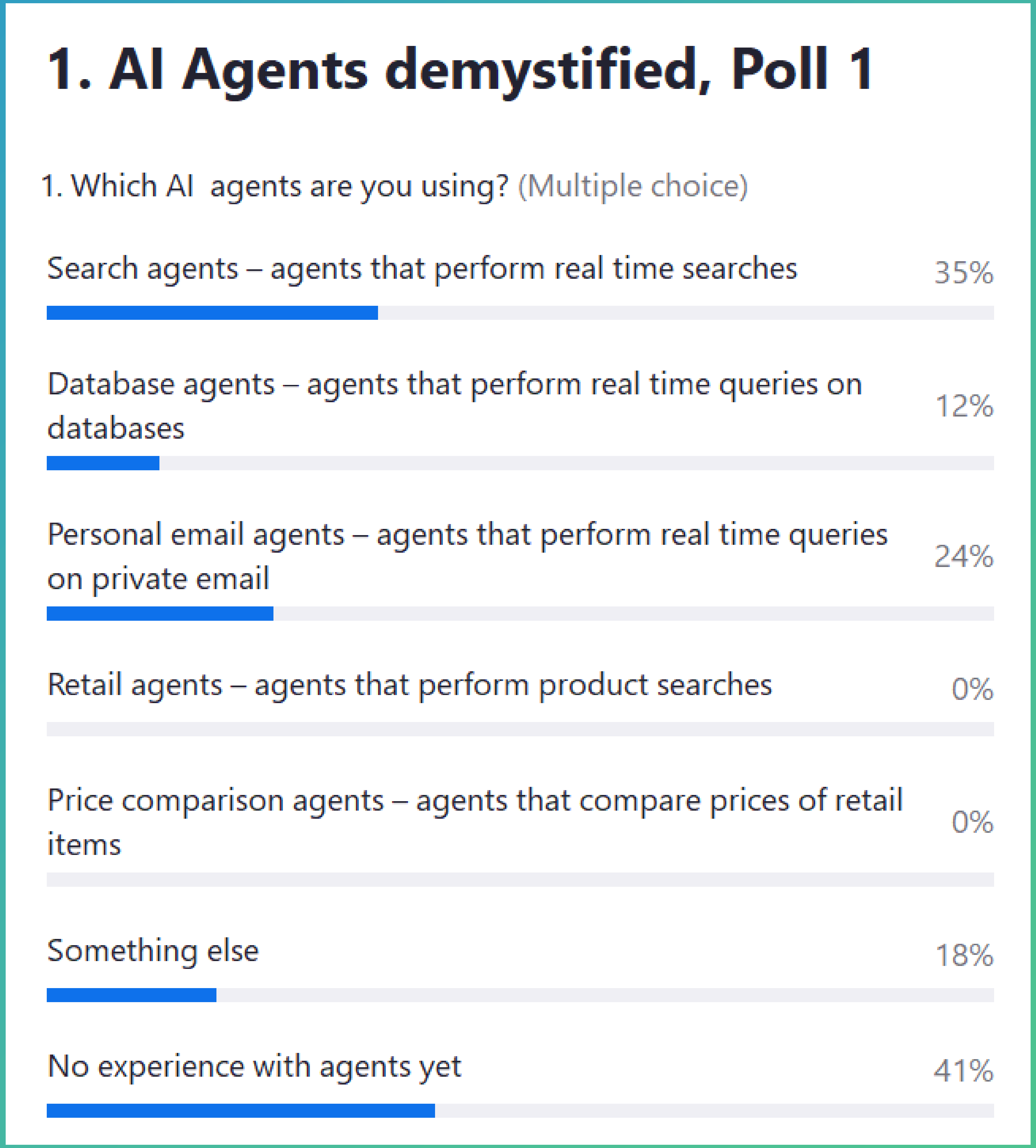
Insights from previous webinar

LLMs suited for many scenarios

- Great at a variety of NLP tasks
- Converting unstructured data into structured data
- Tackling the problem of “dark data”
- Learning from context

- Limited reasoning
- Knowledge cut-offs
- Long-term planning
- Missing sources
- Hallucinations
- Refusals
- **Cannot “execute” some tasks on their own**

Audience insights



1. AI Agents demystified, Poll 1

1. Which AI agents are you using? (Multiple choice)

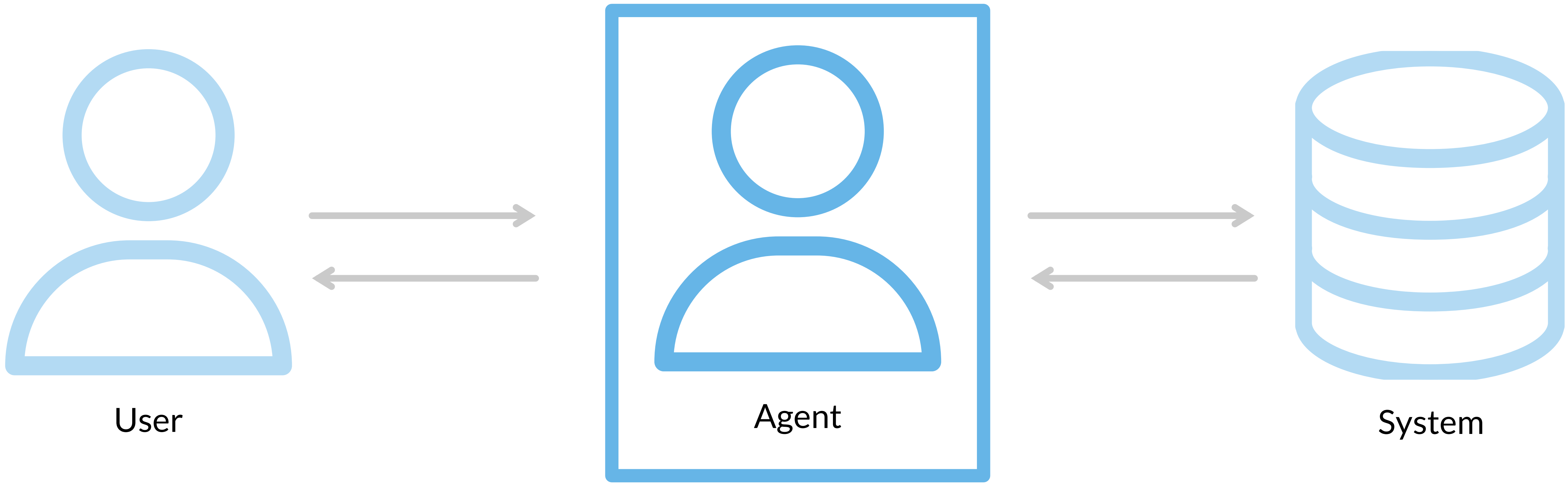
Search agents – agents that perform real time searches	35%
Database agents – agents that perform real time queries on databases	12%
Personal email agents – agents that perform real time queries on private email	24%
Retail agents – agents that perform product searches	0%
Price comparison agents – agents that compare prices of retail items	0%
Something else	18%
No experience with agents yet	41%

What are agents?

What are agents?

The intermediary

- An agent acts like an intermediary between a user and a system.
- An analogy would be the real estate agent or the insurance broker.



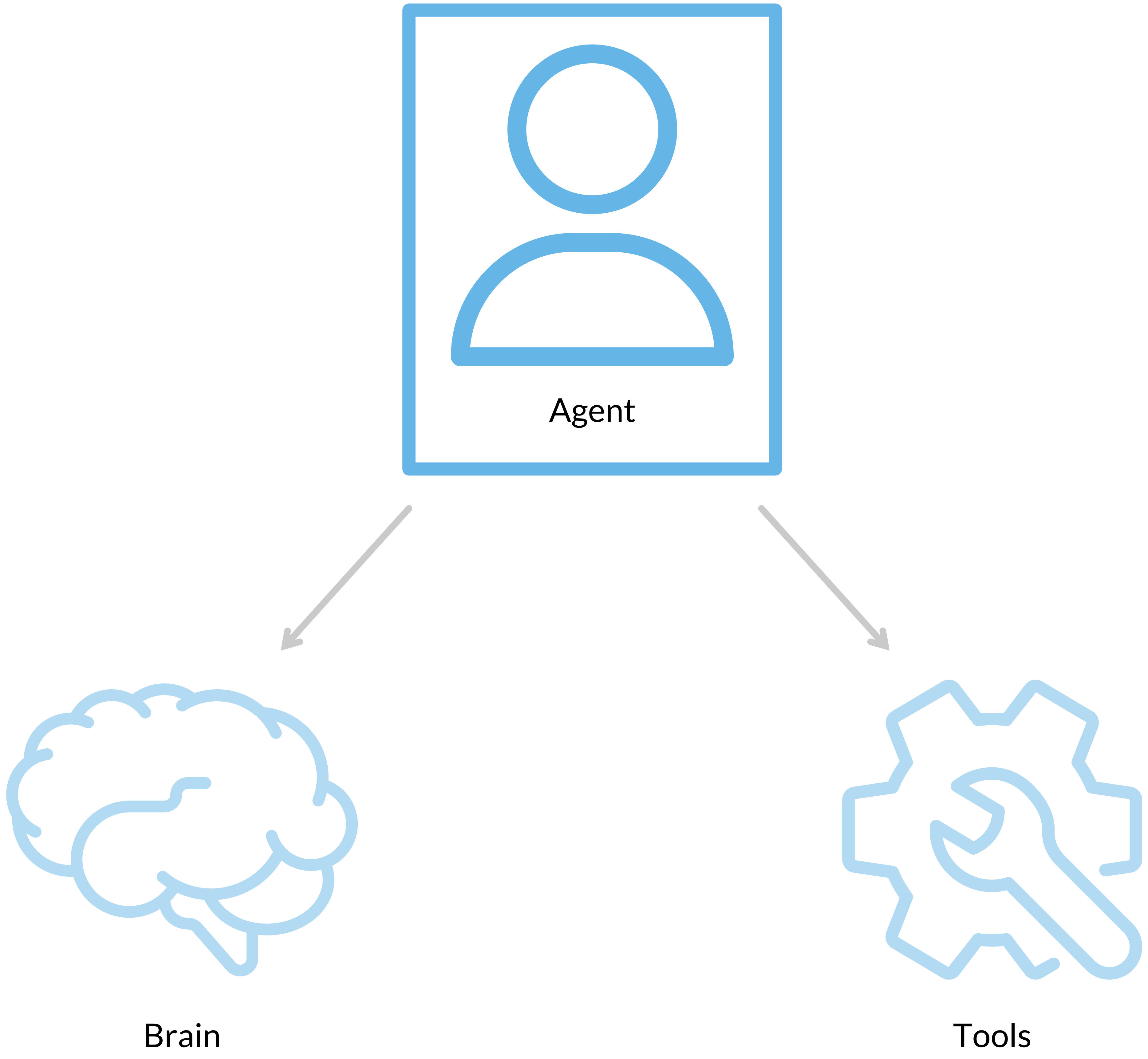
What are agents?

Tools and brain

The agent has two faculties:

Brain:
the central intelligence part which takes the decision.

Tools:
Use the right tools to produce the final answer.



Brain – Intelligence and Memory

The brain has two parts:

Intelligence

The brain has a reasoning part and also the capability to remember.

Memory

Memory is crucial for an agent to work, so that it can create a normal conversation.

Memory

There are different types of memory that are relevant for user agents:

- Chat history
- External knowledge base
- User profile
- Internal caching – previous interactions

Non-deterministic execution path

Deterministic

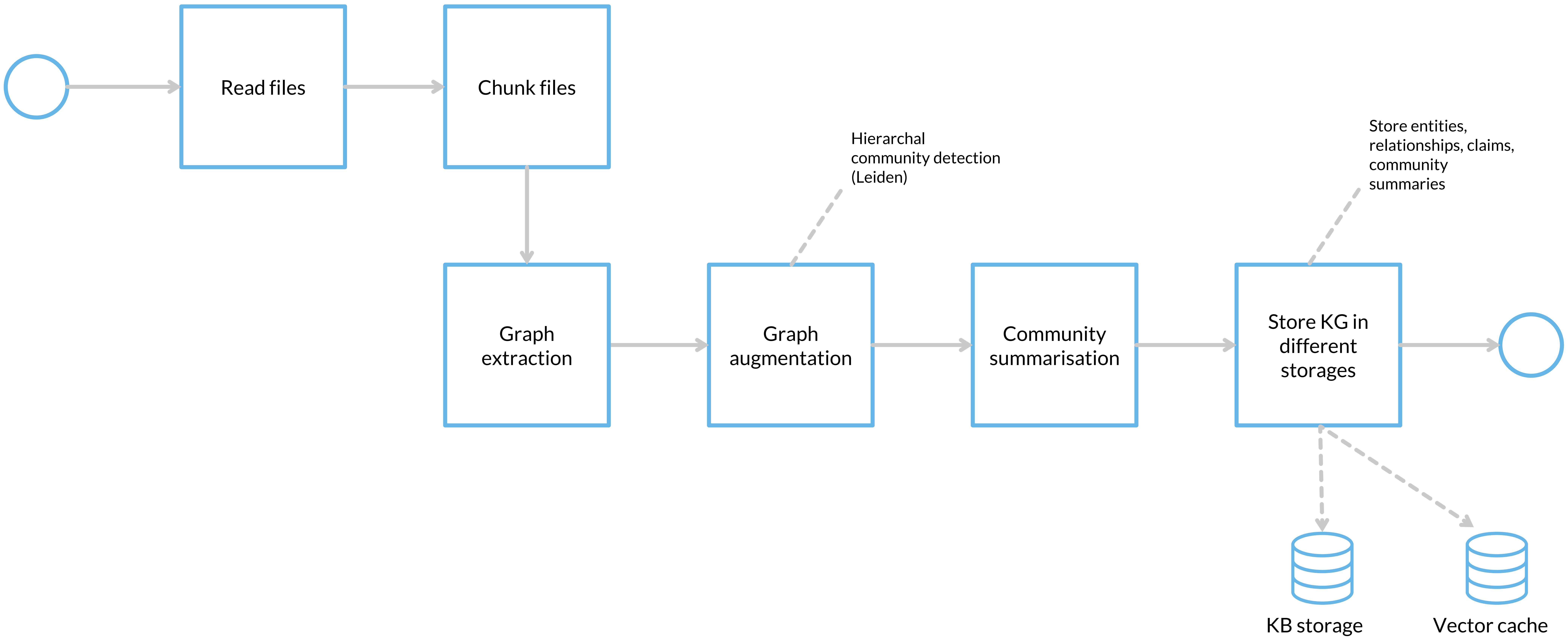
- AI Workflows have a fixed or pre-determined execution path.
- They might succeed or fail, but not flexible.

Non-deterministic

- The agent has an execution path that can vary. The order or execution of the tasks is not pre-determined.
- They are flexible and unpredictable.

What are agents?

Deterministic execution path (example)



Read files

Chunk files

Graph extraction

Graph augmentation

Hierarchal community detection (Leiden)

Community summarisation

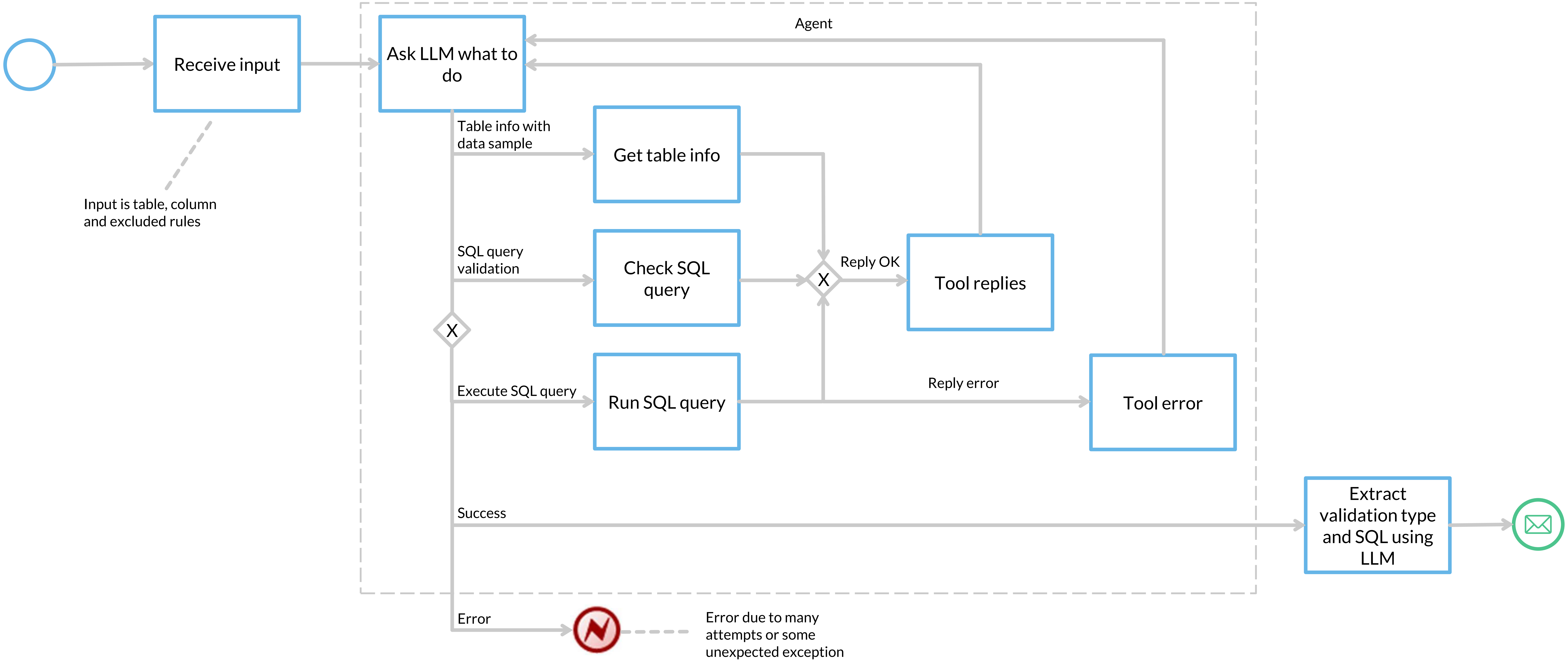
Store KG in different storages

Store entities, relationships, claims, community summaries

KB storage

Vector cache

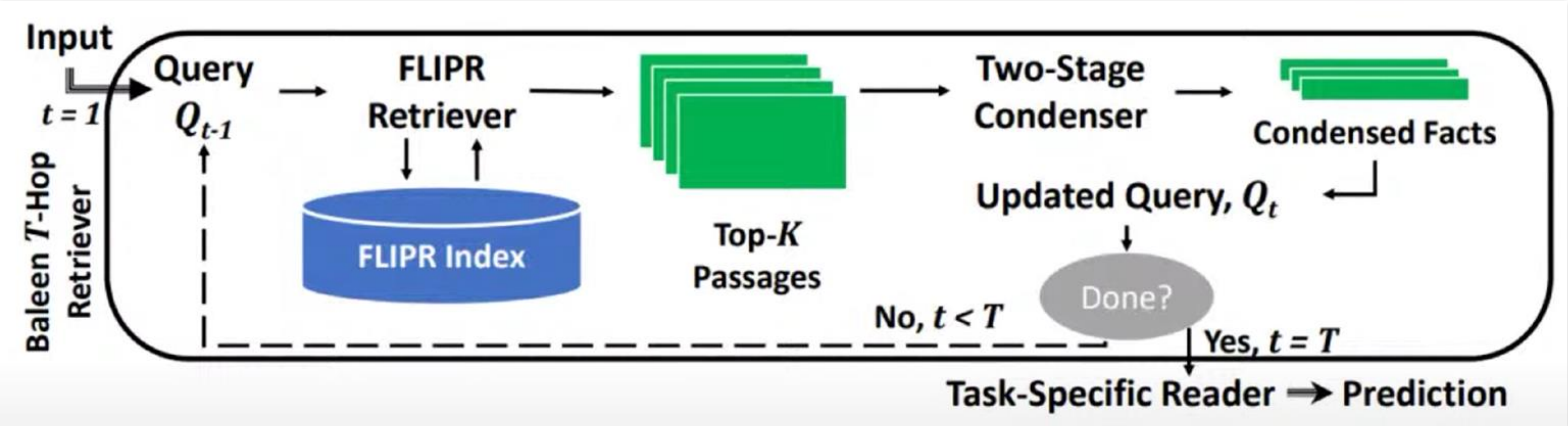
Non-deterministic execution path (example)



From RAG to agents

Multi-Hop RAG

- Ask LLM to break down questions in pieces
- Find information about each piece
- Returns information about each piece to the LLM
- LLM synthetises (summarises) response for each
- Uses smart techniques to map passages to partial questions and the order of retrieval of information
- After retrieval and condensation, system compiles the final answer



Multi-Hop RAG

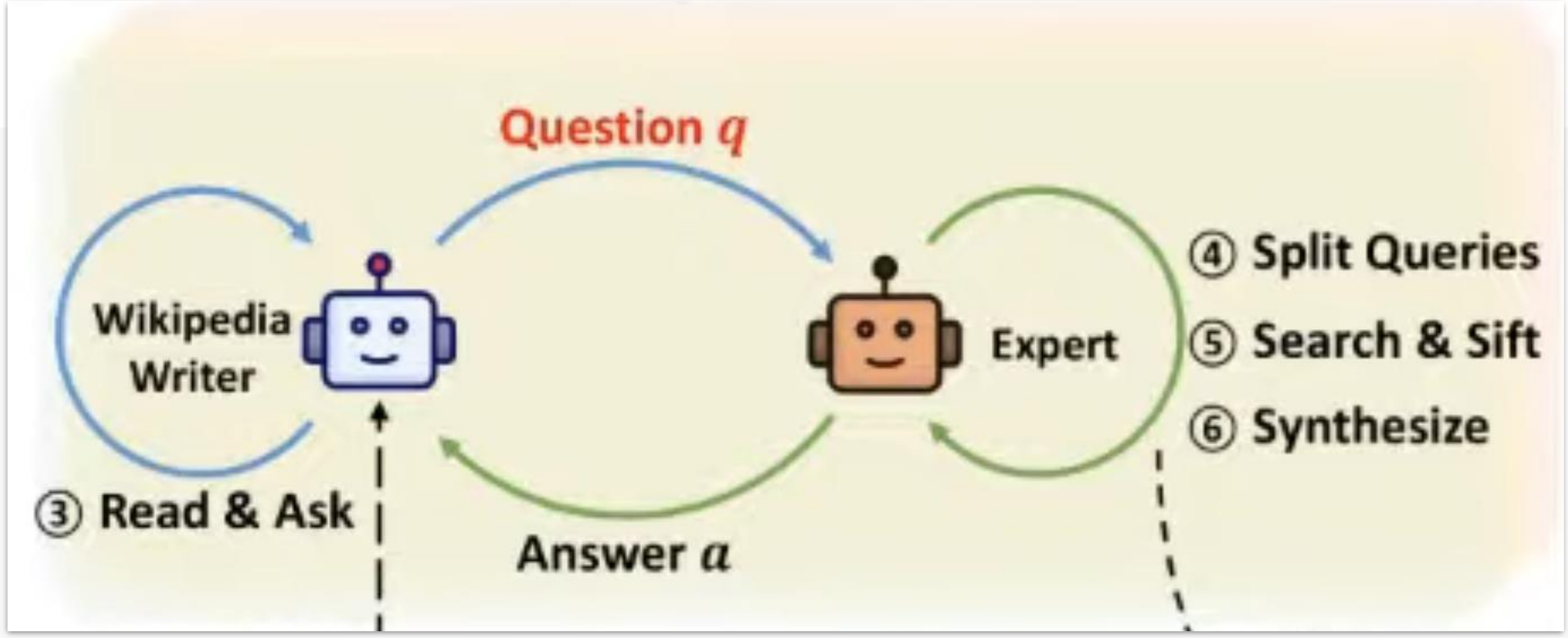
Reasoning across multiple documents

Source: github.com/stanford-futuredata/Baleen?tab=readme-ov-file

From RAG to Agents

STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking)

- STORM is a RAG system that uses two agents, a Wikipedia Writer and an expert to develop a topic.
- Agents are able to improve the quality of the generation by interacting with each other.



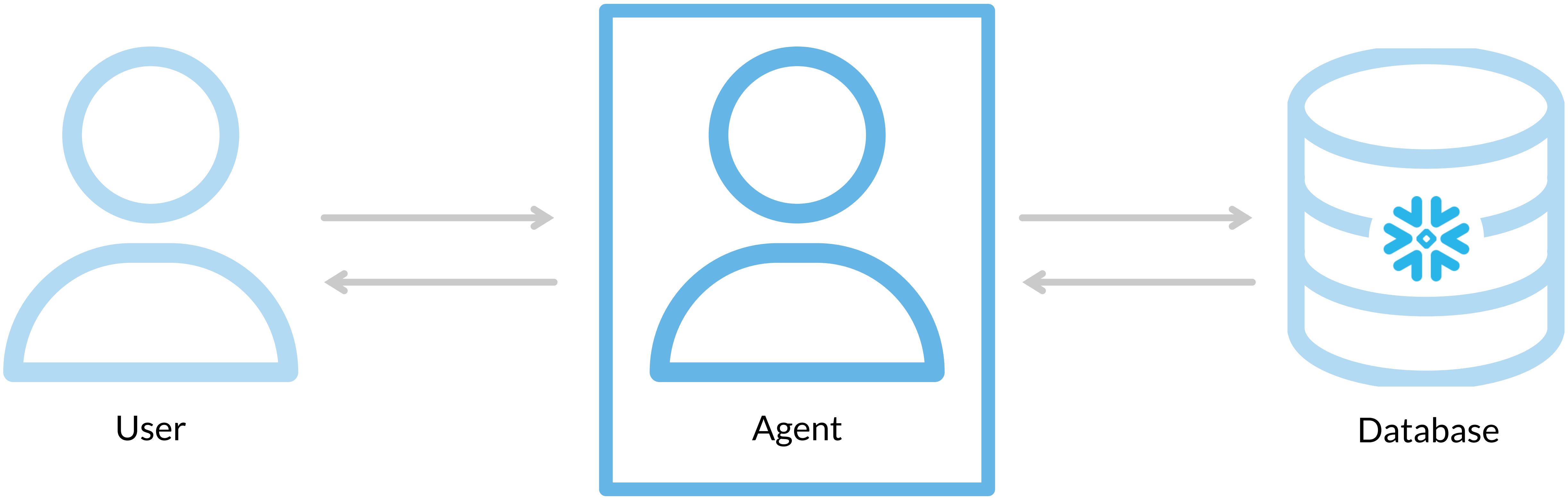
Source: github.com/stanford-futuredata/Baleen?tab=readme-ov-file

Agent examples

Agent examples

Examples – SQL Agent

Intermediary that translates questions in nature language to SQL queries which Snowflake understands with the goal to retrieve insights in natural language to the user.

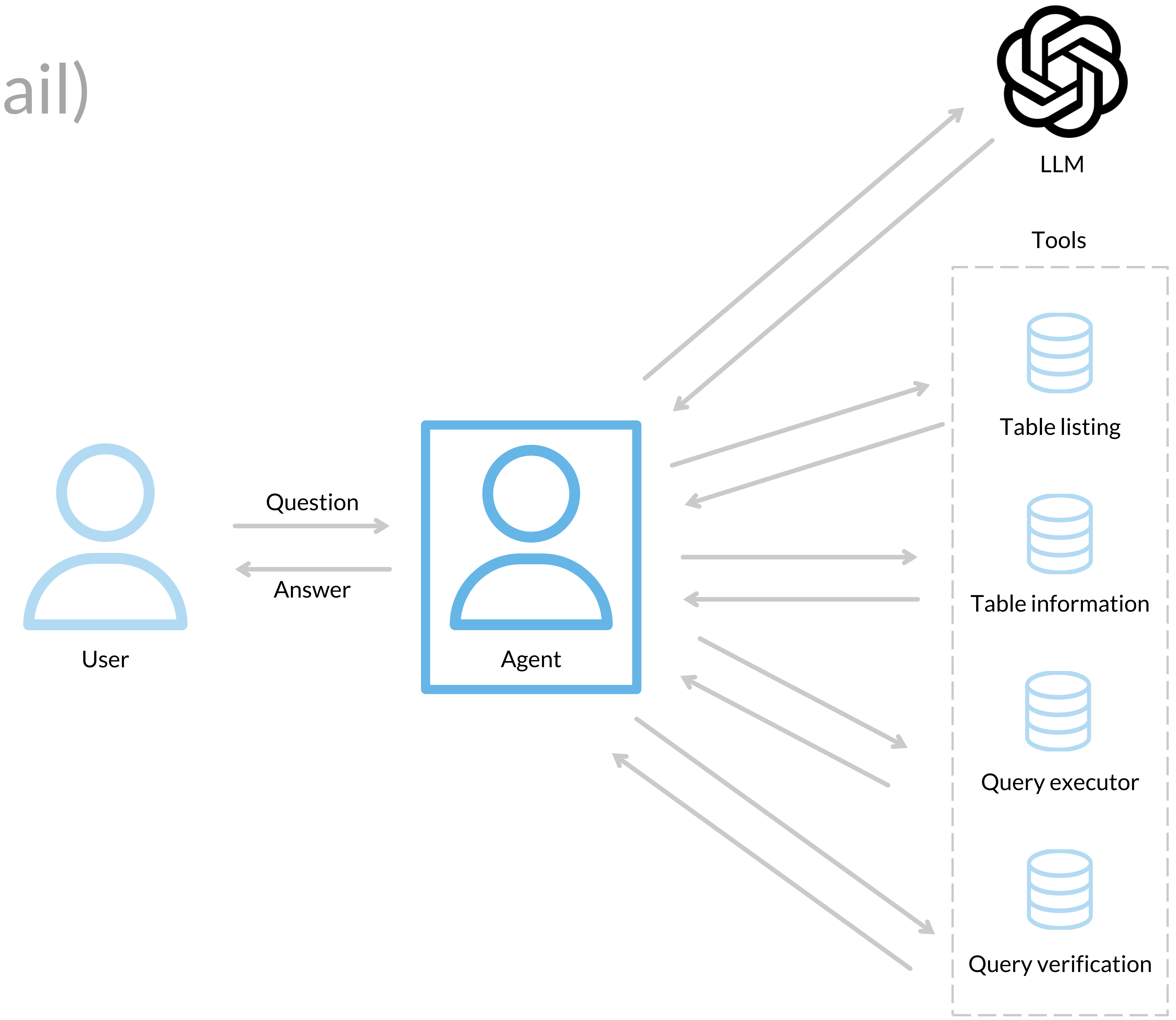


Agent examples

Examples – SQL Agent (Detail)

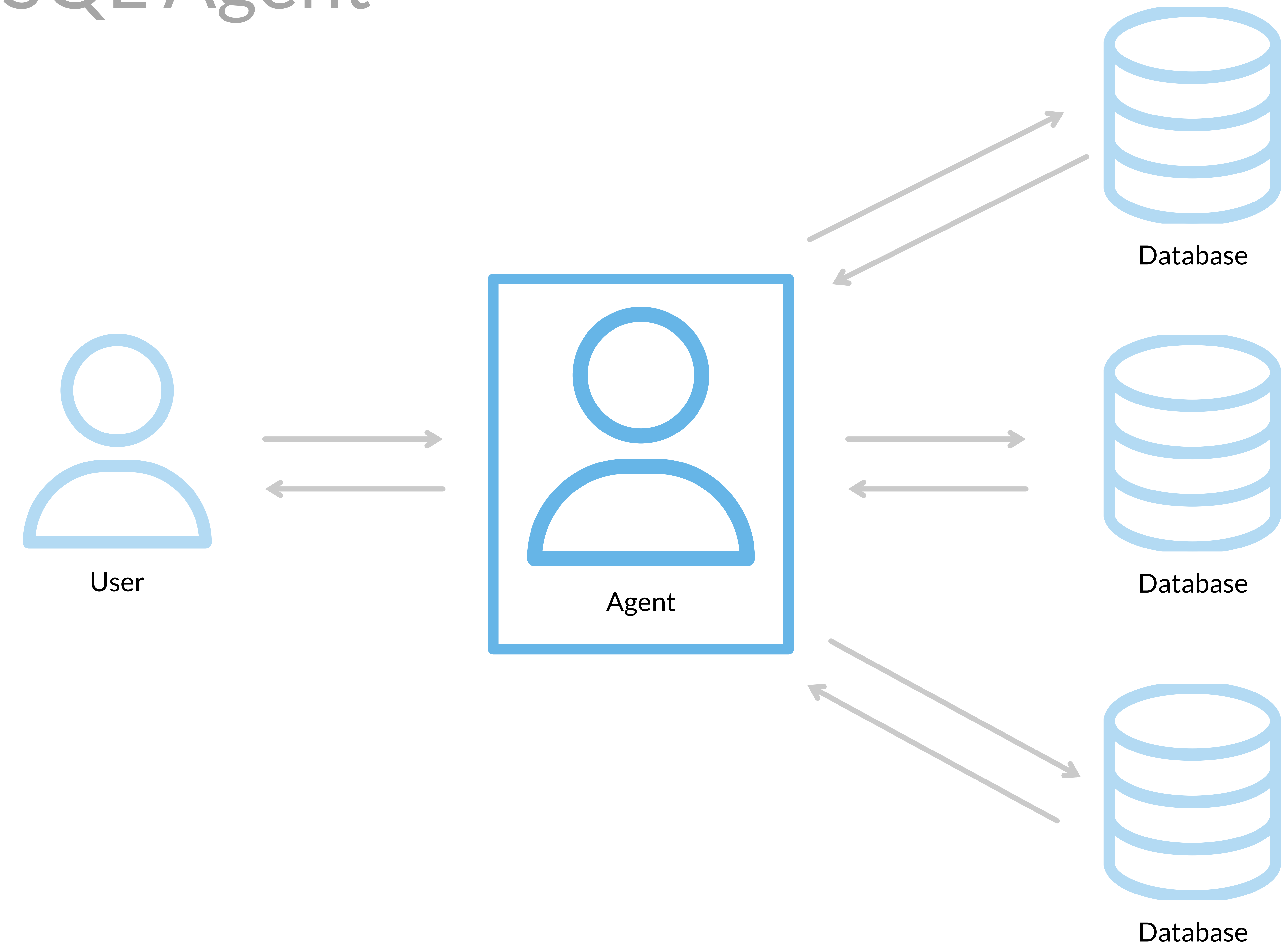
In reality the agent has multiple tools:

- Table listing tool
- Table information tool
- Query executor tool
- Query verification tool



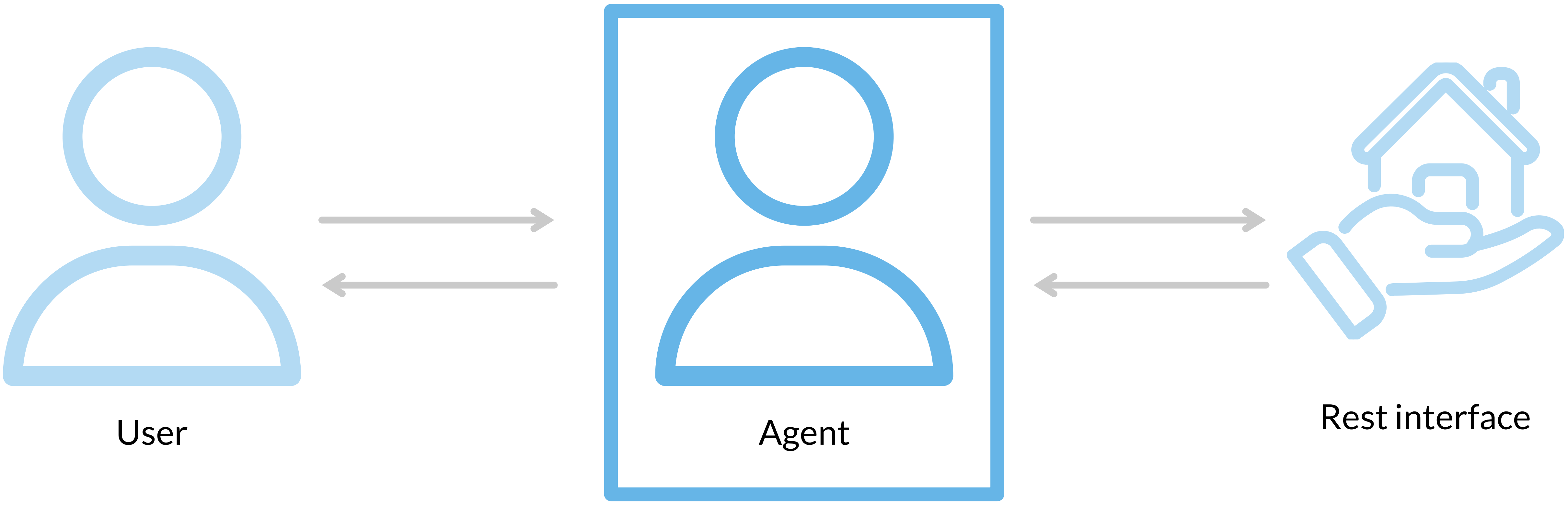
Examples – Multi-database SQL Agent

Intermediary that communicates with multiple databases.



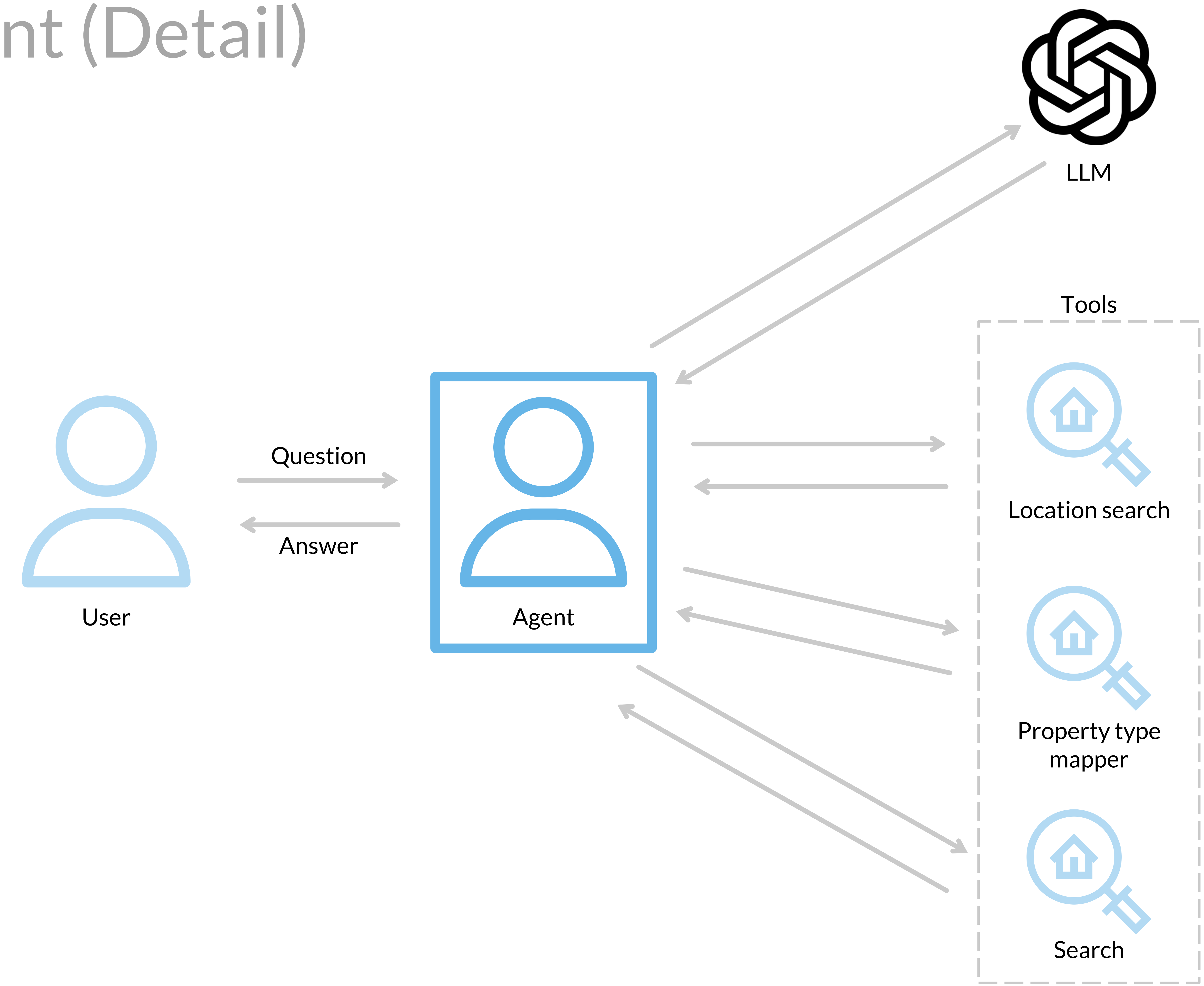
Examples – Real Estate Agent

Intermediary that translates questions in natural language to queries against a REST or similar interface which retrieves housing properties.



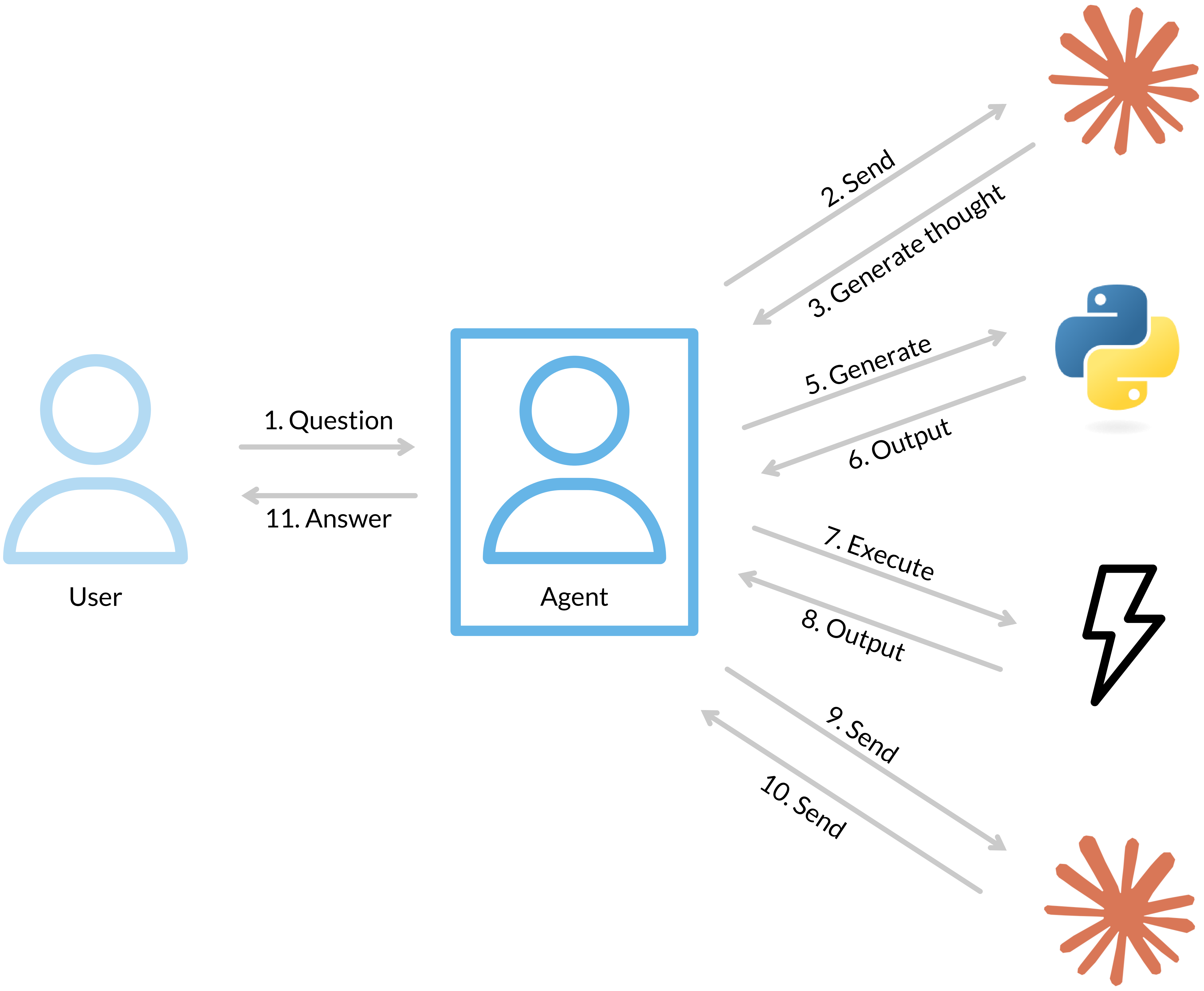
Examples – Real Estate Agent (Detail)

In practical terms you will have multiple tools that access different REST interfaces.



Examples – Programme of Thought

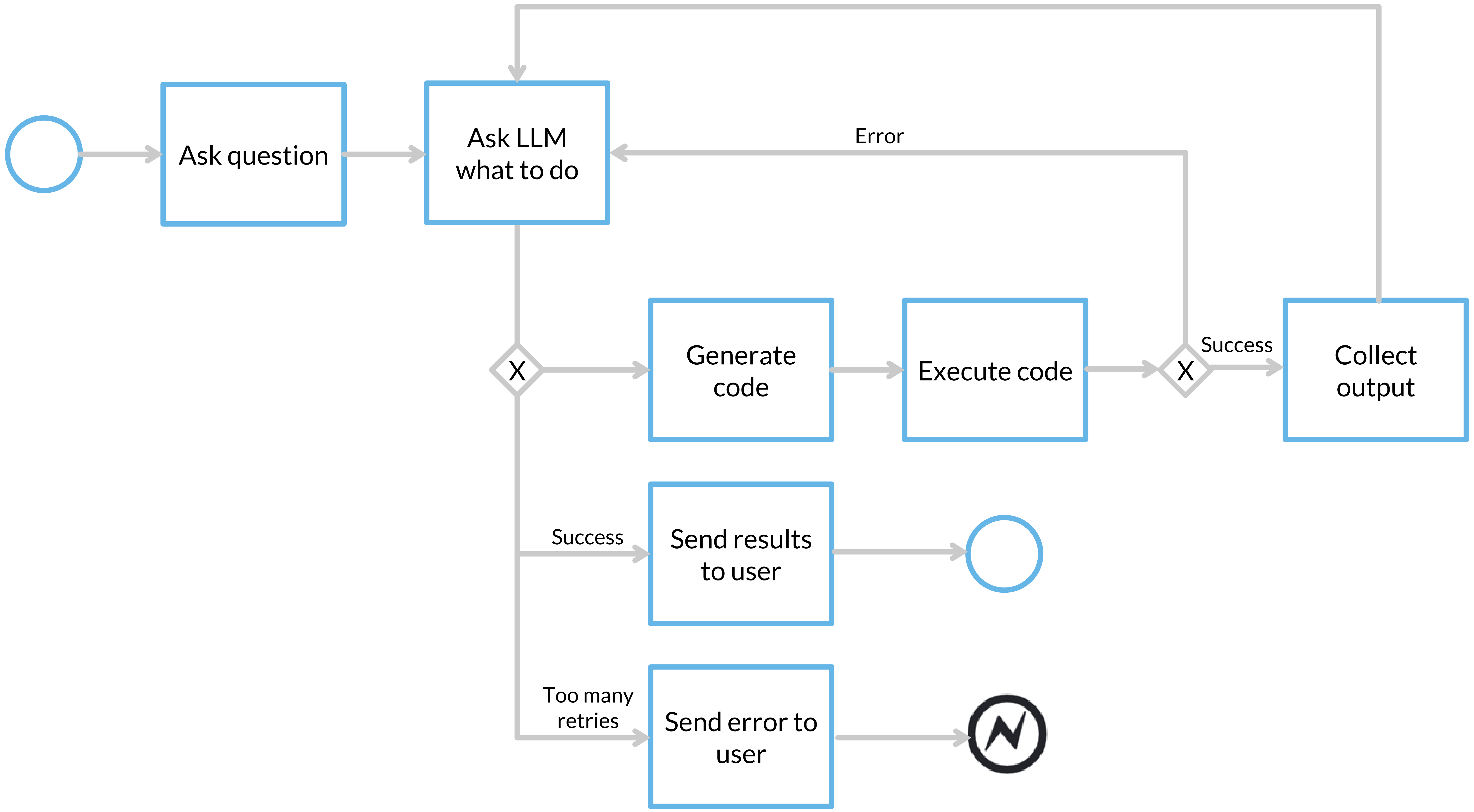
Intermediary that translates questions in natural language to a software programme. The software programme is then executed, and its results are then used to formulate an answer.



Agent examples

Examples – Programme of thought (Detail)

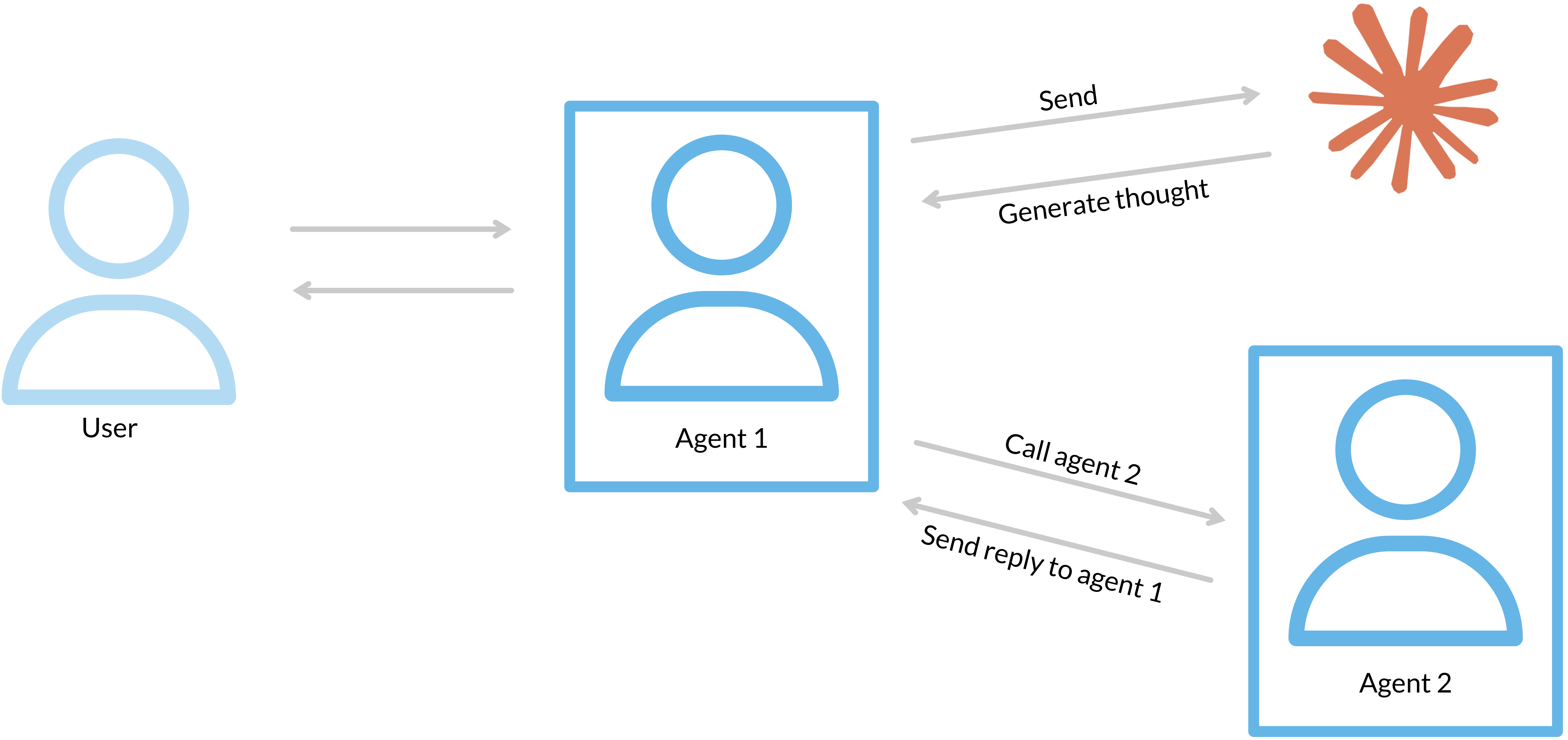
In practice program of thought consists of a workflow which tries to ensure that potential errors are also handled.



Demo

Examples – Agent with agents as tools

More complex agents have other agents as tools.



ReAct
(Reason and Act)

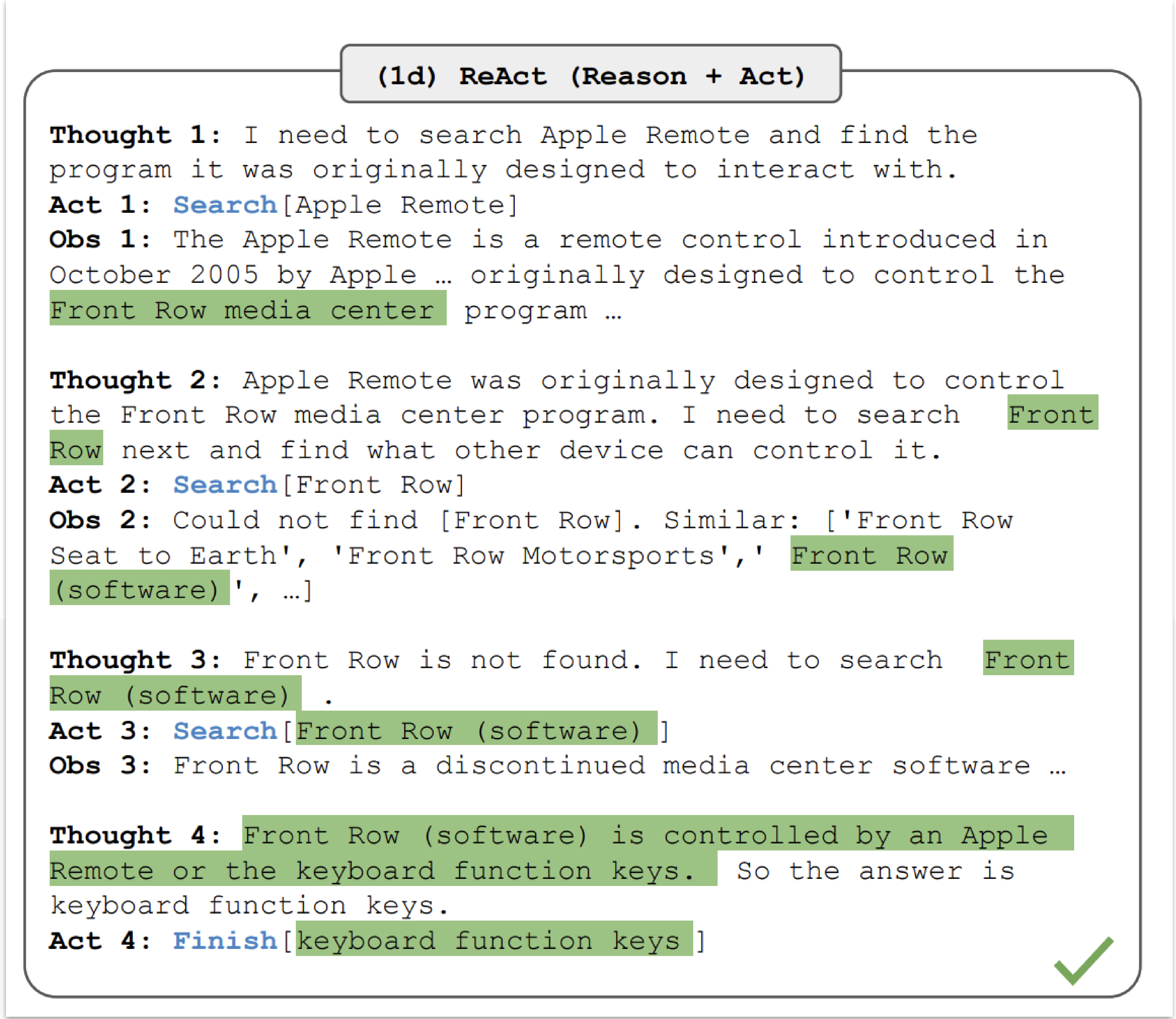
ReAct

Reason and Act

Foundational workflow pattern used in agents

Allows to perform multi-step reasoning tasks

Allows to access information not available to an LLM



(1d) ReAct (Reason + Act)

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]

✓

Source: arxiv.org/pdf/2210.03629

ReAct

Main ingredients

Who	Verb	Function
LLM	Think	Create context
LLM	Act	Decide tools to call
Tool	Observe	Call function and get result

Interaction example

Question	Which are the best sites from where you can get AI related news? Can you include the URLs please?
Reason	I should look up the best sites for AI-related news to provide accurate and up-to-date URLs. I will perform a search using the **serp_api_tool** to get a list of reliable sources for AI news.
Act	Function: serp_api_tool Arguments: “best sites for AI news with URLs”
Observe	JSON with search results: https://www.enago.com/academy/guestposts/carlcroger/ai-news-websites/ • AIChief — Your Ultimate Companion For AI News • IEEE Spectrum — A Legit Source of AI Industry News • TechTarget — An AI News Website To Get Insides. …
Answer	Here are some of the best sites for AI-related news along with their URLs: …

Typical ReAct base prompts

User prompt example

Which are the most sold cars in April 2024?

Internal Prompt 1

- You will be given {inputs_} and your goal is to finish with {outputs_}.
- **To do this, you will interleave Thought, Tool Name, and Tool Args, and receive a resulting Observation.**
- Thought can reason about the current situation, and Tool Name can be the following types:

  <tool1> <tool2> …

Typical ReAct base prompts

User prompt example

What will the weather be like tomorrow?

Internal Prompt 2

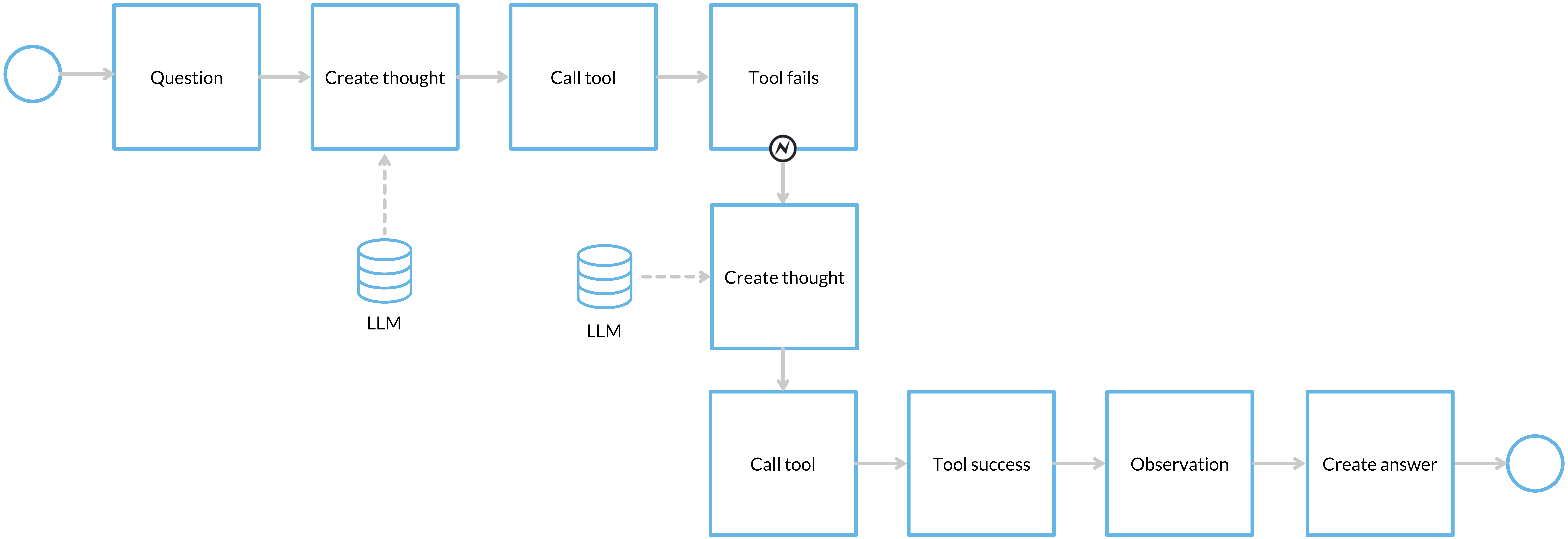
- **You run in a loop of Thought, Tool, PAUSE, Observation.**
- At the end of the loop you output an Answer, when you have the final answer. Make sure you include "Answer:" in your output in this case.
- Use Thought to describe your thoughts about the question you have been asked.
- Use Tool to run one of the tools available to you - then return PAUSE. Make sure you include exactly the term "PAUSE" in your output in this case.
- Observation will be the result of running those tools.

Self improving agents

Self improving agents

Error correction

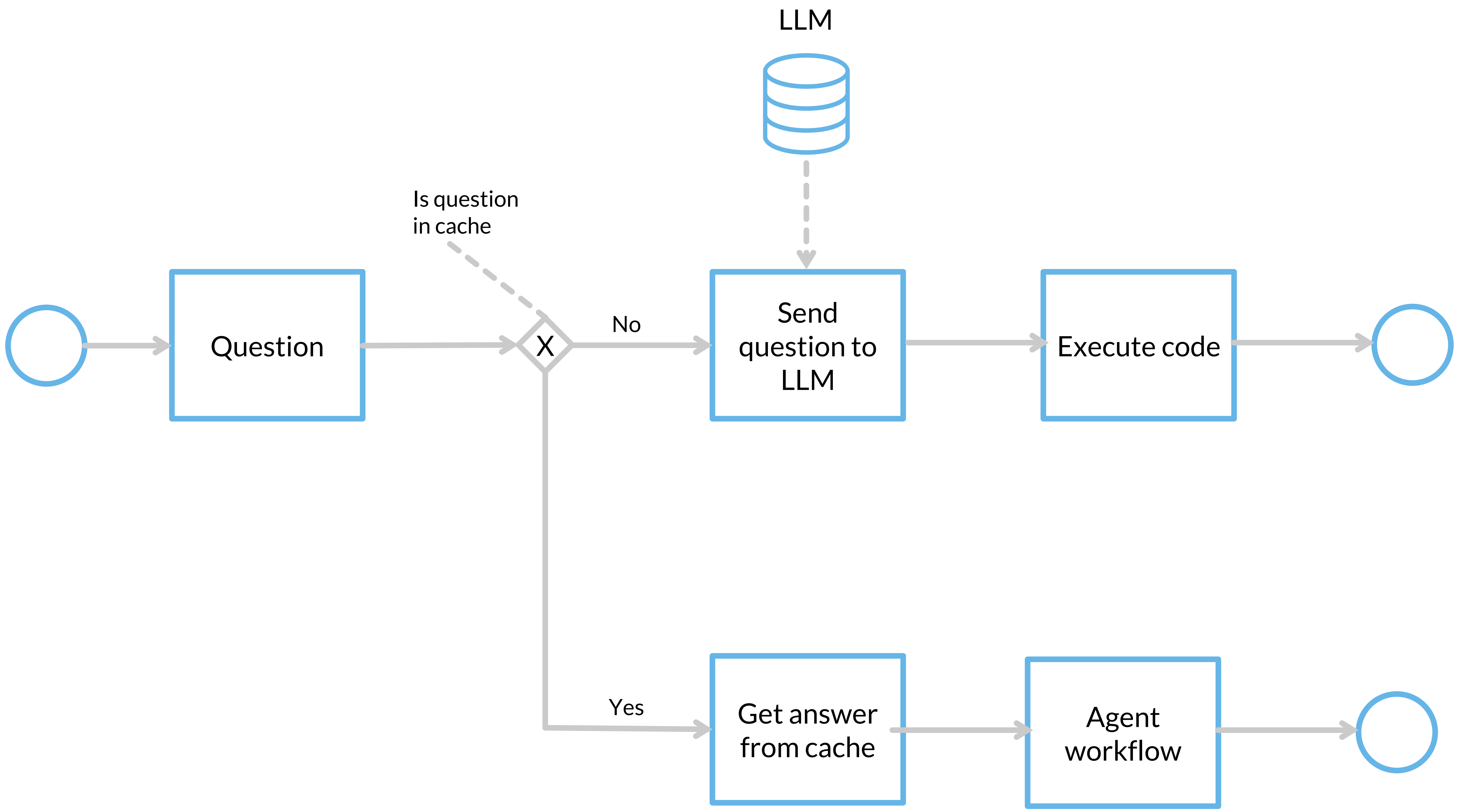
Agents can deal with errors and retry its execution an undefined number of times.



Self improving agents

Caching

Agents should be able to cache results to avoid extra costs and improve speed.



Auto Prompt Tuning

Main idea:

instead of creating prompts manually
you use a dataset with training samples
and an optimizer to fine tune the prompt.

Auto Prompt Tuning

Prompt Signature

Example:
- Query
- Possible outputs
- Result

Dataset

Fields:
- Query
- Possible outputs
- Result

Prompt Optimizer

Examples:
- Sample examples from dataset (multi shot)
- Use LLM to generate examples based on the dataset (multi shot)

Evaluator

Components:
- Metrics

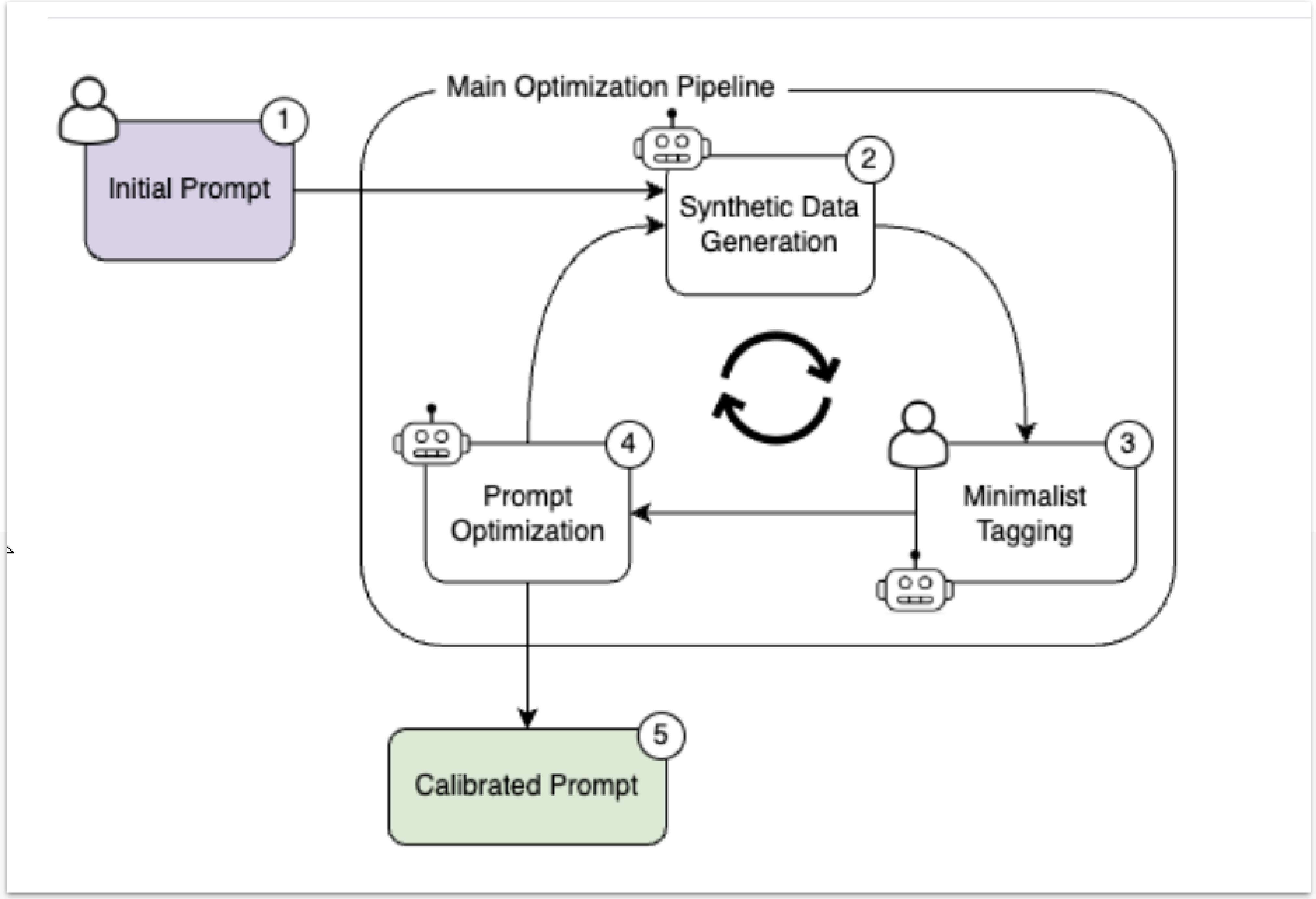
Self improving agents

Auto prompt tuning frameworks

DSPy

Declarative Self-improving Python

Source: https://dspy.ai/



AutoPrompt

Source: github.com/Eladlev/AutoPrompt

Audience poll

In which areas do you think you can use agents?

A. Product search (e.g. Retail)
B. Product recommendation (e.g. Retail)
C. Gathering information (e.g. Consulting)
D. Generating recommendations (e.g. Consulting)
E. Reporting (e.g. Consulting)

E. Document search (e.g. Legal firms)
F. Trouble shooting (e.g. helpdesk)
G. Candidate search (e.g. Human Resources)
H. Database search (IT related)

Agent toolkits

Self improving agents

Agent toolkits

Semantic Kernel

AutoGen

DSPy

LangGraph

crewai

What’s next?

Self improving agents

What’s next?

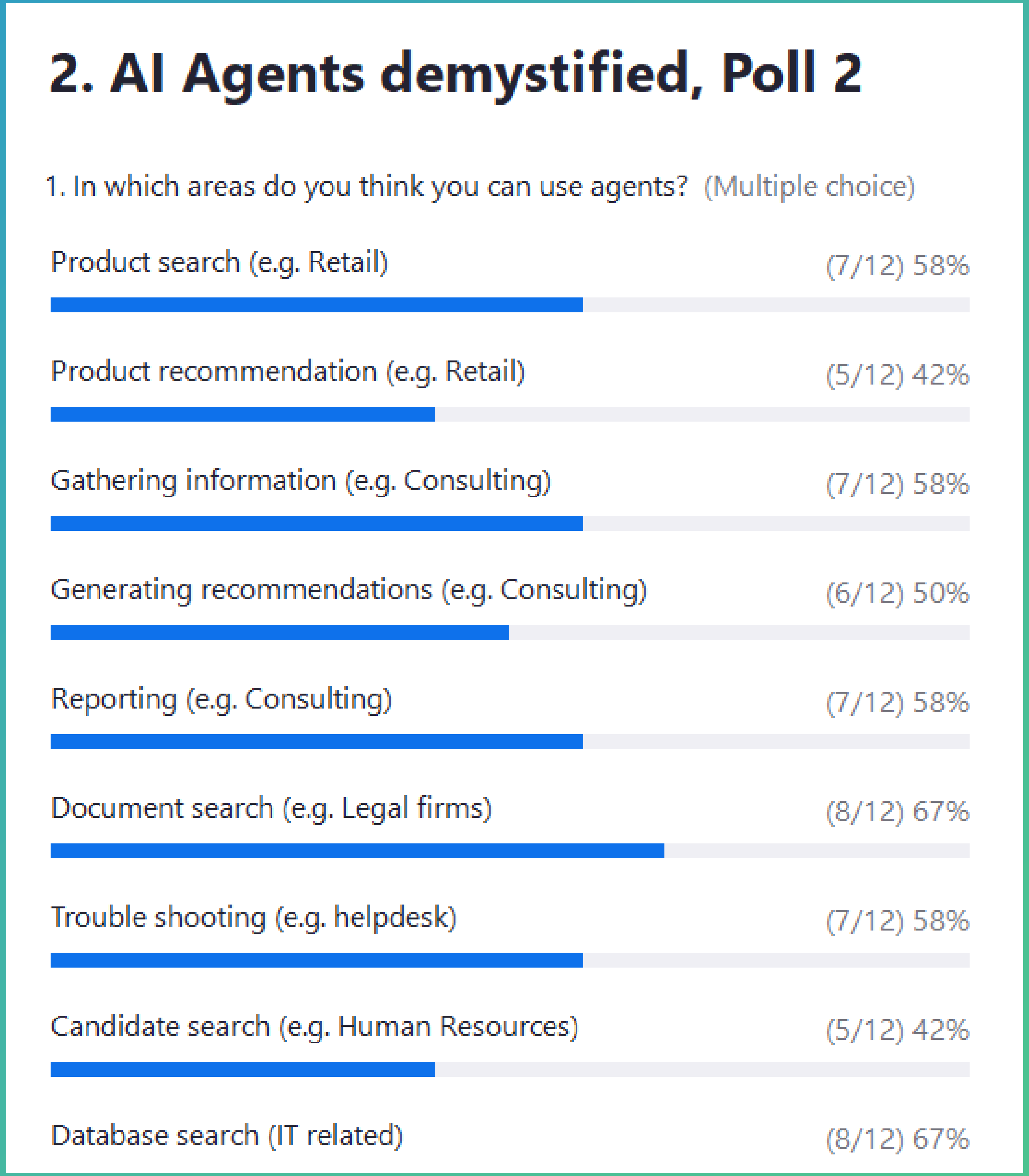
Voice and video agents

Agents with ears and mouth which understand spoken language and can also speak

Agent swarms

Teams of agents that work towards a common goal

Audience insights



2. AI Agents demystified, Poll 2

1. In which areas do you think you can use agents? (Multiple choice)

Product search (e.g. Retail)	(7/12) 58%
Product recommendation (e.g. Retail)	(5/12) 42%
Gathering information (e.g. Consulting)	(7/12) 58%
Generating recommendations (e.g. Consulting)	(6/12) 50%
Reporting (e.g. Consulting)	(7/12) 58%
Document search (e.g. Legal firms)	(8/12) 67%
Trouble shooting (e.g. helpdesk)	(7/12) 58%
Candidate search (e.g. Human Resources)	(5/12) 42%
Database search (IT related)	(8/12) 67%

Credits

chatgpt.com

DSPy

dspy.ai

AutoGen

microsoft.github.io/autogen/0.2

LangGraph

github.com/langchain-ai/langgraph

LangChain

langchain-ai.github.io/langgraph/tutorials/rag/langgraph_agentic_rag

Semantic Kernel

learn.microsoft.com/en-us/semantic-kernel/overview

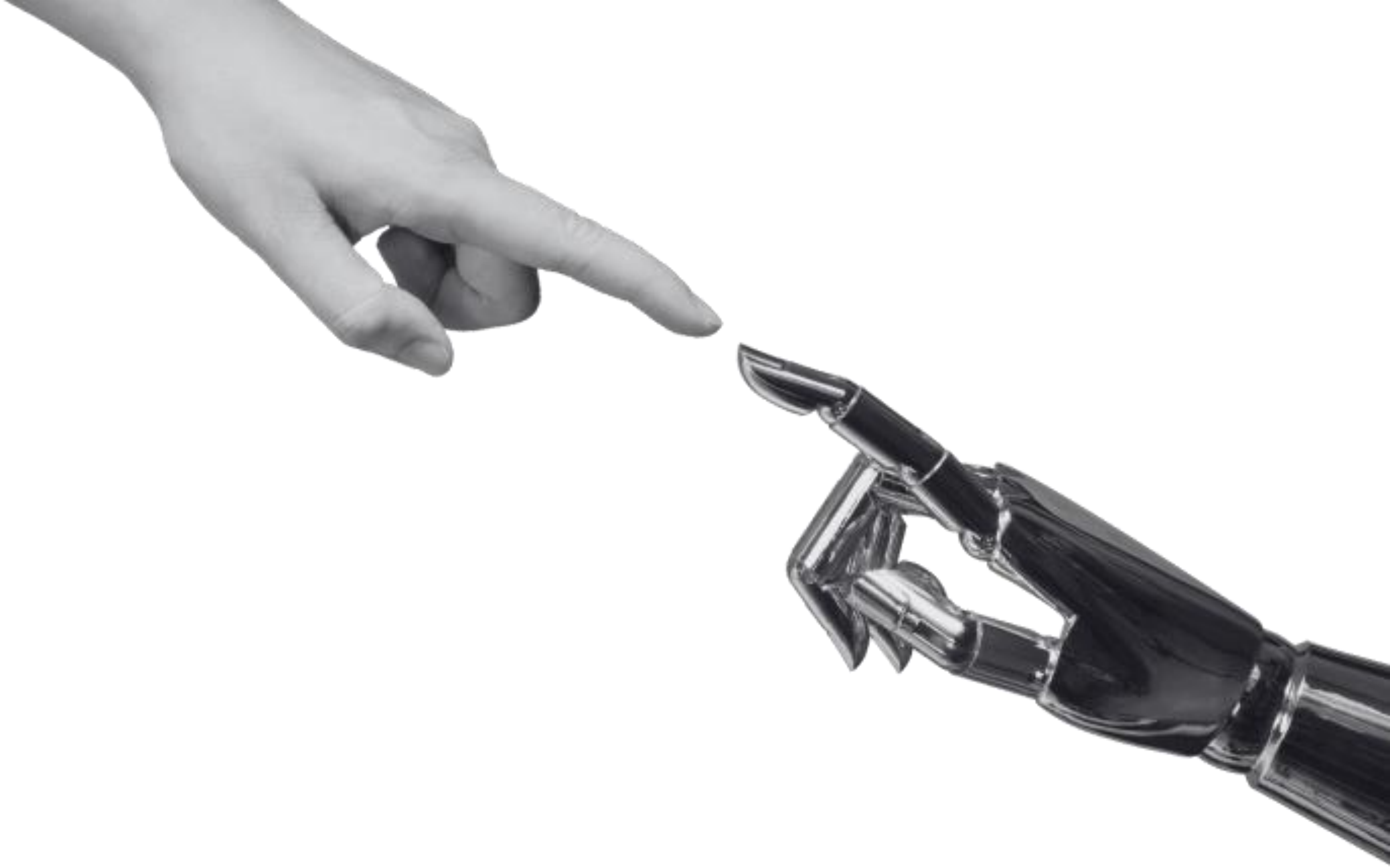
crewai

crewai.com

REACT:
Synergizing Reasoning and Acting in Language Models

arxiv.org/pdf/2210.03629

Thank you for joining



Please feel free to contact Gil Fernandes if you have any feedback about the session.

Email at techtalk@onepointltd.com

Connect on LinkedIn www.linkedin.com/in/gil-palma-fernandes

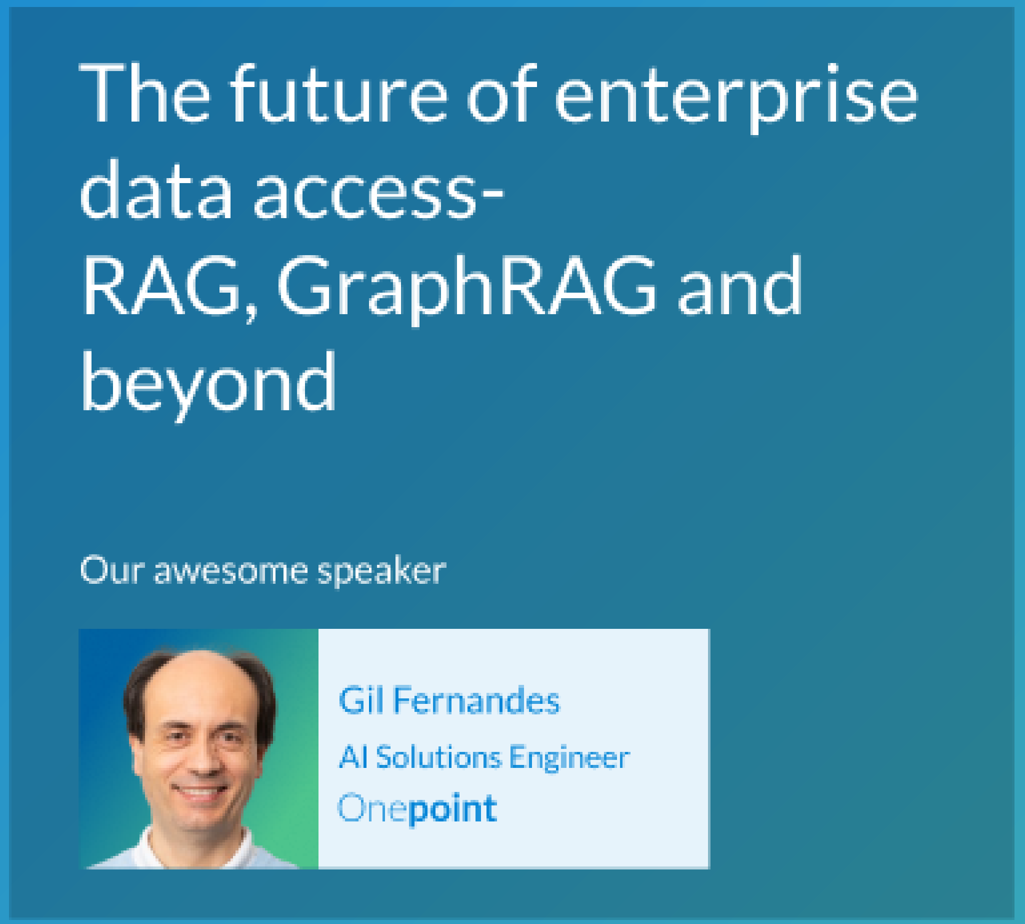
Find Gil’s Reflections on AI at medium.com/@gil.fernandes

Onepoint
TechTalk

Missed previous webinars?

Watch the replays at onepointltd.com/techtalk

Recent webinars



The future of enterprise data access-
RAG, GraphRAG and beyond

Our awesome speaker

Gil Fernandes
AI Solutions Engineer
Onepoint

How to turn the spotlight on Dark Data with Gen AI

Our awesome speaker

Allan Schweitz
Director of Technology - Services
Onepoint

AI and Master Data Management:
A synergy for success

Our awesome speakers

Dr Madassar Manzoor
Principal Data Architect
Onepoint

Tom Clarke
Data Management Lead
boomi