



2024 AI+研发数字峰会

AI+ Development Digital summit

AI驱动研发迈进数智化时代

中国·上海 05/17-18

AI Agent: 认知框架与案例实践

黄佳 新加坡科技研究局

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **K+峰会**  **上海站**
K+ 全球软件研发行业创新峰会
时间: 2024.06.21-22

 **K+峰会**  **敦煌站**
K+ 思考周®研习社
时间: 2024.10.17-19

 **K+峰会**  **香港站**
K+ 思考周®研习社
时间: 2024.11.10-12



K+峰会详情



 **AiDD峰会**  **上海站**
AI+研发数字峰会
时间: 2024.05.17-18

 **AiDD峰会**  **北京站**
AI+研发数字峰会
时间: 2024.08.16-17

 **AiDD峰会**  **深圳站**
AI+研发数字峰会
时间: 2024.11.08-09



AiDD峰会详情



黄佳

AI研究员 《动手做AI Agent》《GPT图解》作者

新加坡科技研究局AI研究员、技术作家，主攻方向为大语言模型的开发和应用实践、AI in FinTech、AI in MedTech、持续学习，代表作主要有《大模型应用开发 动手做AI Agent》《GPT图解》《零基础学机器学习》《数据分析咖哥十话》《LangChain实战课》等，曾在埃森哲新加坡公司担任多年SAP技术顾问，负责政府和企业薪酬系统的实施和运维。

目录

CONTENTS

1. 何为AI Agent
2. Agent认知框架
3. 适用场景分析
4. 具体实现/技术实践
5. 总结与展望

PART 01

何为AI Agent

▶ 生成式人工智能应用5层级

表 1 生成式人工智能应用需要经历的 5 个层级

层级	AI 应用	描述	示例
L1	Tool (工具)	人类完成所有工作, 没有任何明显的 AI 辅助	Excel、Photoshop、MATLAB 和 AutoCAD 等绝大多数应用
L2	Chatbot (聊天机器人)	人类直接完成绝大部分工作。人类向 AI 询问, 了解信息。AI 提供信息和建议但不直接处理工作	初代 ChatGPT
L3	Copilot (协同)	人类和 AI 共同工作, 工作量相当。AI 根据人类要求完成工作初稿, 人类进行后期校正, 修改和调整, 并最终确认	GitHub Copilot, Microsoft Copilot
L4	Agent (智能体)	AI 完成绝大部分工作, 人类负责设定目标、提供资源和监督结果, 以及最终决策。AI 完成任务拆分, 工具选择, 进度控制, 实现目标后自主结束工作	AutoGPT, BabyAGI, MetaGPT
L5	Intelligence (智能)	完全无需人类监督, AI 自主拆解目标、寻找资源, 选择并使用工具, 完成全部工作, 人类只需给出初始目标	类冯·诺伊曼机器人或者 人?

▶ 什么是Agent

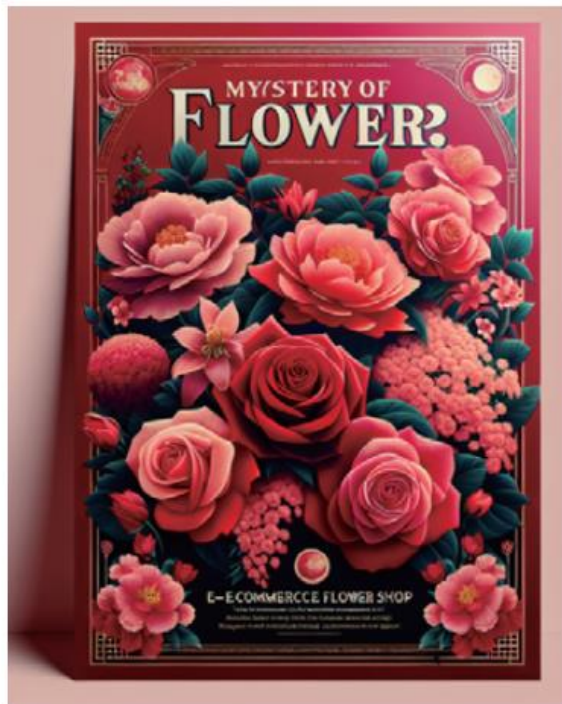


Agent的核心组件

▶▶ Agent和传统程序设计范式究竟有何不同?

- 进行推理 CoT, Reflection
- 调用工具 ReAct, Function Calling/Tool Calls
- 自我进化

进行推理



一张鲜花海报

```
> Entering new AgentExecutor chain...
  我应该使用图片创作说明文案来描述这张图片
Action: Image captioner
Action Input: https://mir-s3-cdn-cf.bahance.net/project_modules/hd/eec79e20058499.563190744f903.jpg
Observation: there is a picture of flowers with the word Love
Thought: 这张图片传达了爱的信息, 我应该强调这一点
Action: Image captioner
Action Input: https://mir-s3-cdn-cf.bahance.net/project_modules/hd/eec79e20058499.563190744f903.jpg
Observation: there is a picture of flowers with the word Love
Thought: 我可以加入一些词语来描述这张图片, 比如美丽、浪漫、温馨
Action: Image captioner
Action Input: https://mir-s3-cdn-cf.bahance.net/project_modules/hd/eec79e20058499.563190744f903.jpg
Observation: there is a picture of flowers with the word Love
Thought: 我可以加入一些情感色彩, 让这张图片更具吸引力
Action: Image captioner
Action Input: https://mir-s3-cdn-cf.bahance.net/project_modules/hd/eec79e20058499.563190744f903.jpg
Observation: there is a picture of flowers with the word Love
Thought: 我现在知道了最终的答案, 我可以写出一篇吸引人的中文推广文案了
Final Answer: 这张美丽的图片传达了爱的信息, 让我们一起分享这份浪漫温馨的感觉吧!

> Finished chain.
```

Agent的自主思考过程

```
> Entering new AgentExecutor chain...
  我应该想一想该怎么做
Action: Image captioner
Action Input: https://mir-s3-cdn-cf.bahance.net/project_modules/hd/eec79e20058499.563190744f903.jpg
Observation: there is a picture of flowers with the word Love
Thought:
  这张图片很美, 我应该把它和爱情联系起来
Action: Image captioner
Action Input: https://mir-s3-cdn-cf.bahance.net/project_modules/hd/eec79e20058499.563190744f903.jpg
Observation: there is a picture of flowers with the word Love
Thought:
  我可以写上"爱情就像这束花, 美丽而持久"
Final Answer: 爱情就像这束花, 美丽而持久

> Finished chain.
```

Agent运行的轮次不同, 思考结果也不同

调用工具

工具集

与知识真实性相关

- 混合搜索
- Web浏览器
- 维基百科
- 视频网站
- 快百科
-

与时间感知能力相关

- 日历
- 节假日
- 时间间隔
- 天气
-

自定义工具

The screenshot shows the OpenAI Playground interface. At the top, there is a status bar with the text "Elevated error rate impacting ChatGPT and API" and a "Status Page" link. The main interface is titled "Playground" and features a sidebar with various icons. The chat area is titled "RAG小能手" and contains the following content:

Name: RAG小能手
asst_yjQ9cE53mId1HkexowyEQcaE

Instructions: 你是我的销售小助理，擅长数据分析

Model: gpt-4-turbo-preview

TOOLS:

- Functions (+ Function)
- Code interpreter (toggle off)
- Retrieval (toggle on)

Chat History:

User: 在鲜花的销售中，我们那个产品销量最好？那个卖的最差，对于卖的差产品，应该采取什么推广策略？

RAG小能手: 为了回答你的问题，我们需要首先分析文档中的销售数据来找出销量最好和最差的产品。然后，我们可以基于这些发现来提出一些推广策略。根据文档中的销售数据，我们首先需要汇总每种鲜花的总销量。我们将计算每种花卖出的总数，然后找出销量最好和最差的鲜花。请稍等片刻，我将为你进行这些计算。

RAG小能手: 在数据分析后，我发现以下销量结果（请注意，由于文档数据量大，以下为基于数据可见部分的初步分析）：

Input field: Enter your message...

Buttons: +, Run

Footer: Playground messages can be viewed by anyone at your organization using the API.

▶▶ 工具使用和模拟是Agent研发的两个重点

- 基于工具的Agent, 以自动化复杂的多步骤工作流程
- 基于模拟的Agent, 目标是在虚拟世界中真实地复制人类行为和社会动态

▶ 模拟



Agent形成的虚拟社会

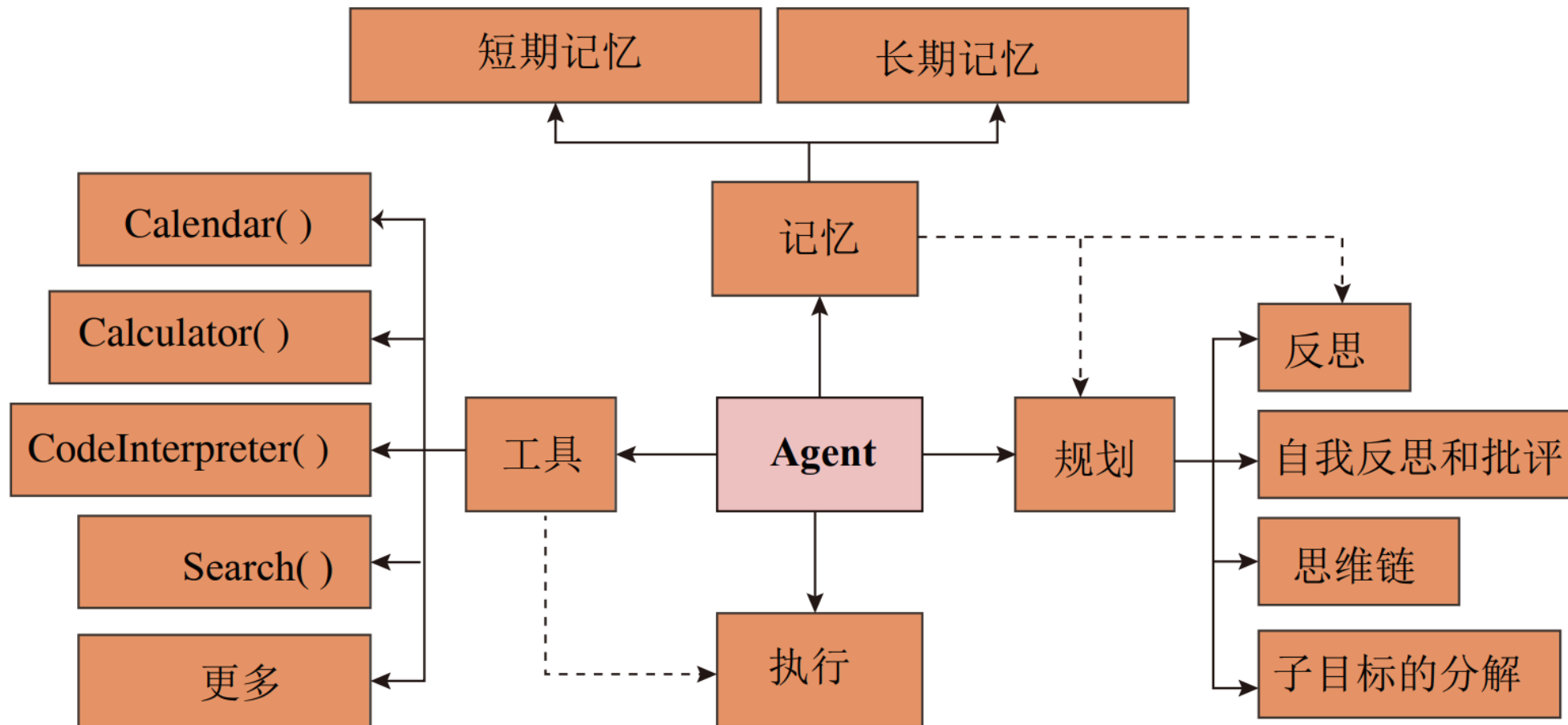
Agent方法论



PART 02

AI Agent认知框架

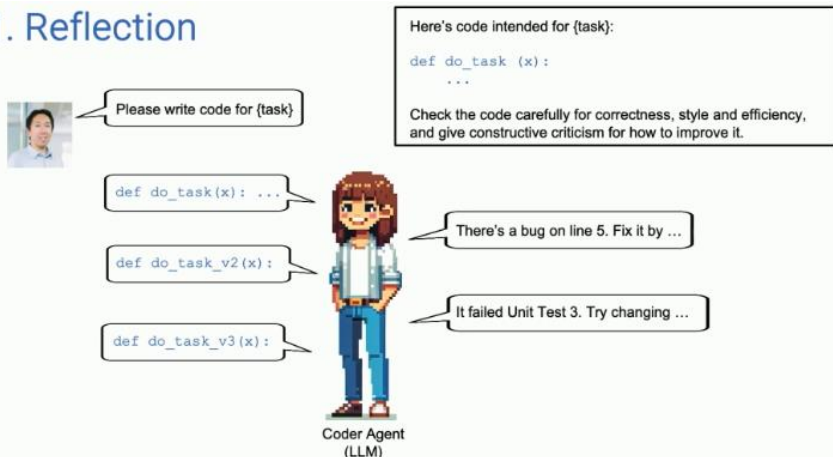
▶ 整体技术框架



由大模型驱动自主Agent的架构（图片来源：LilianWeng博客）

Andrew Ng 给出的四种Agent思维框架设计模式

1. Reflection

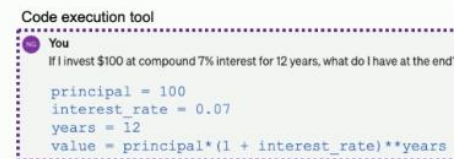


Recommended reading:

- Self-Refine: Iterative Refinement with Self-Feedback, Madaan et al. (2023)
- Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al., (2023)

Andrew Ng

2. Tool use



Analysis

- Code Execution
- Wolfram Alpha
- Bearly Code Interpreter

Research

- Search engine
- Web browsing
- Wikipedia

Productivity

- Email
- Calendar
- Cloud Storage

Images

- Image generation (e.g., Dall-E)
- Image captioning
- Object detection

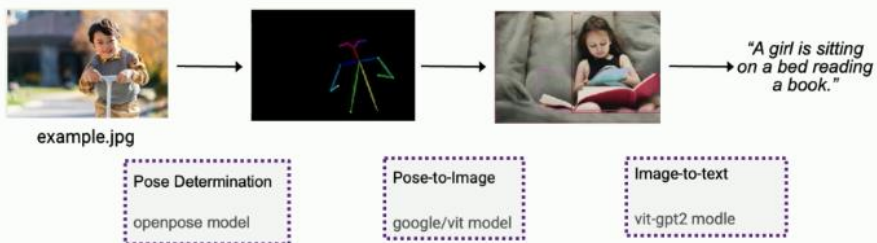
Recommended reading:

- Gorilla: Large Language Model Connected with Massive APIs, Patil et al. (2023)
- MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action, Yang et al. (2023)

Andrew Ng

3. Planning

Request: Please generate an image where a girl is reading a book, and her pose is the same as the boy in the image example.jpg, then please describe the new image with your voice.



[Example adapted from HuggingGPT paper]

Recommended reading:

- Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al., (2022)
- HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, Shen et al. (2023)

Andrew Ng

4. Multiagent collaboration



Multiagent Debate

Task	Single agent	Multi-agent
Biographies	66.0%	73.8%
MMLU	63.9%	71.1%
Chess move	29.3%	45.2%

(Du et al., 2023)

Recommended reading:

- Communicative Agents for Software Development, Qian et al., (2023)
- AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, Wu et al. (2023)

Andrew Ng

▶ Agent认知框架: CoT (Chain of Thought)

Standard Prompting

Model 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: 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?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model 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?

Model Output

A: 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. ✅

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models ,
36th Conference on Neural Information Processing Systems (NeurIPS 2022).

▶ Agent认知框架: Self-Ask

Direct Prompting

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Harry Vaughan Watkins.

Question: Who was president of the U.S. when superconductivity was discovered?
Answer: Franklin D. Roosevelt ❌

Chain of Thought

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died.
So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Who was president of the U.S. when superconductivity was discovered?
Answer: Superconductivity was discovered in 1911 by Heike Kamerlingh Onnes. Woodrow Wilson was president of the United States from 1913 to 1921. So the final answer (the name of the president) is: Woodrow Wilson. ❌

Self-Ask

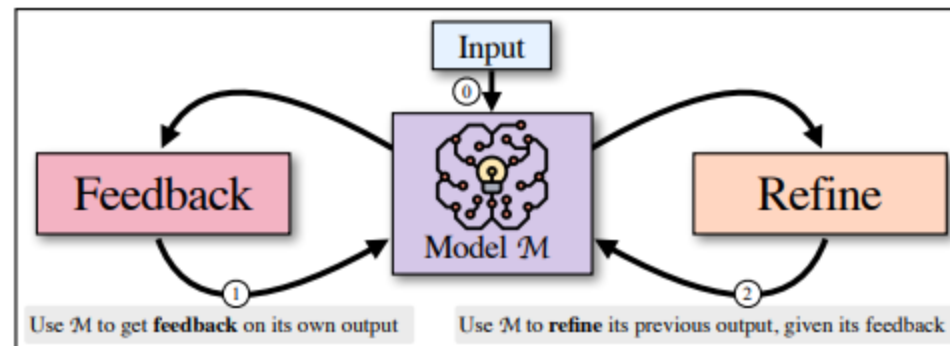
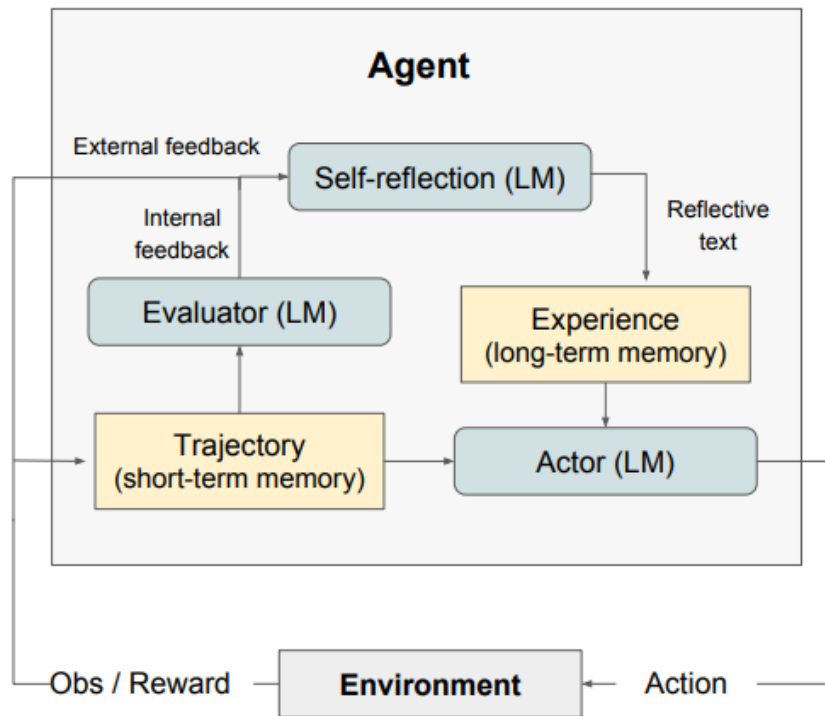
GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Are follow up questions needed here: Yes.
Follow up: How old was Theodor Haecker when he died?
Intermediate answer: Theodor Haecker was 65 years old when he died.
Follow up: How old was Harry Vaughan Watkins when he died?
Intermediate answer: Harry Vaughan Watkins was 69 years old when he died.
So the final answer is: Harry Vaughan Watkins

Question: Who was president of the U.S. when superconductivity was discovered?
Are follow up questions needed here: Yes.
Follow up: When was superconductivity discovered?
Intermediate answer: Superconductivity was discovered in 1911.
Follow up: Who was president of the U.S. in 1911?
Intermediate answer: William Howard Taft.
So the final answer is: William Howard Taft. ✅

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., & Lewis, M. (2022). Measuring and Narrowing the Compositionality Gap in Language Models. arXiv preprint arXiv:2212.09551.

Agent认知框架: Reflection



Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv preprint arXiv:2303.11366.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., ... & Clark, P. (2023). SELF-REFINE: Iterative Refinement with Self-Feedback. arXiv preprint arXiv:2303.17651.

▶ Agent认知框架：Function Callings/Tool Calls

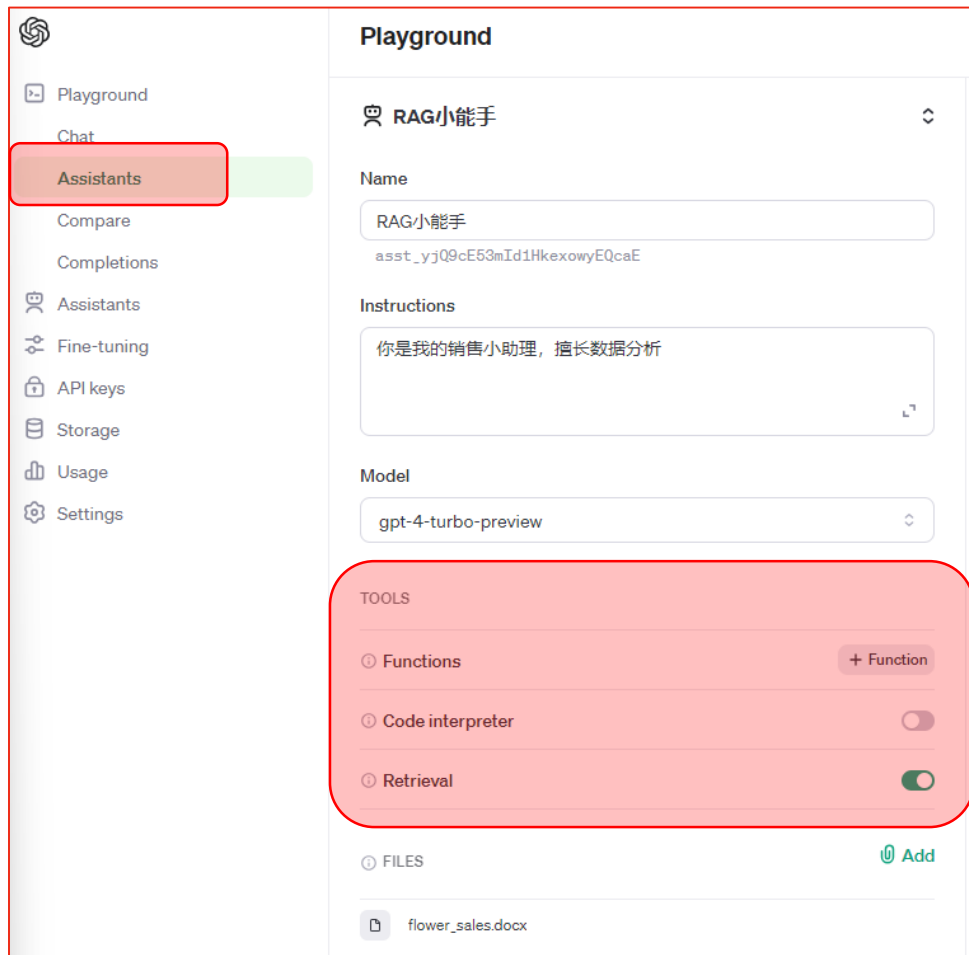
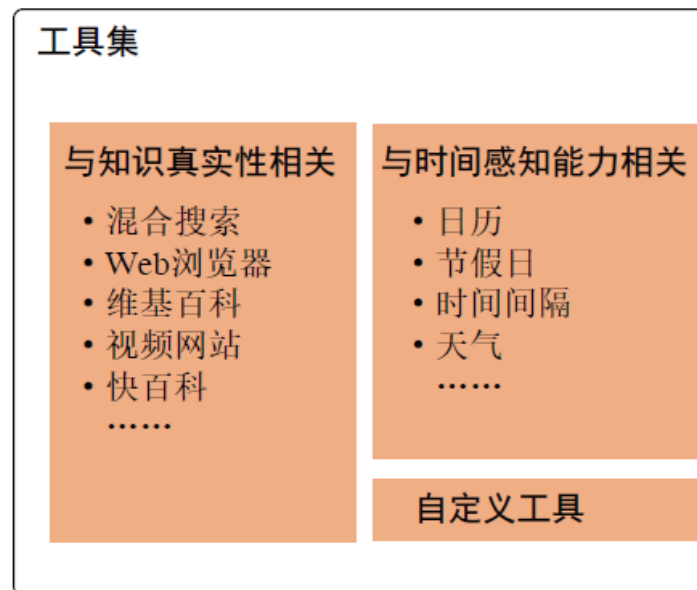
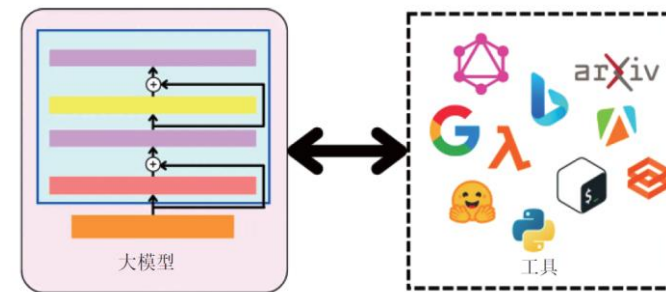
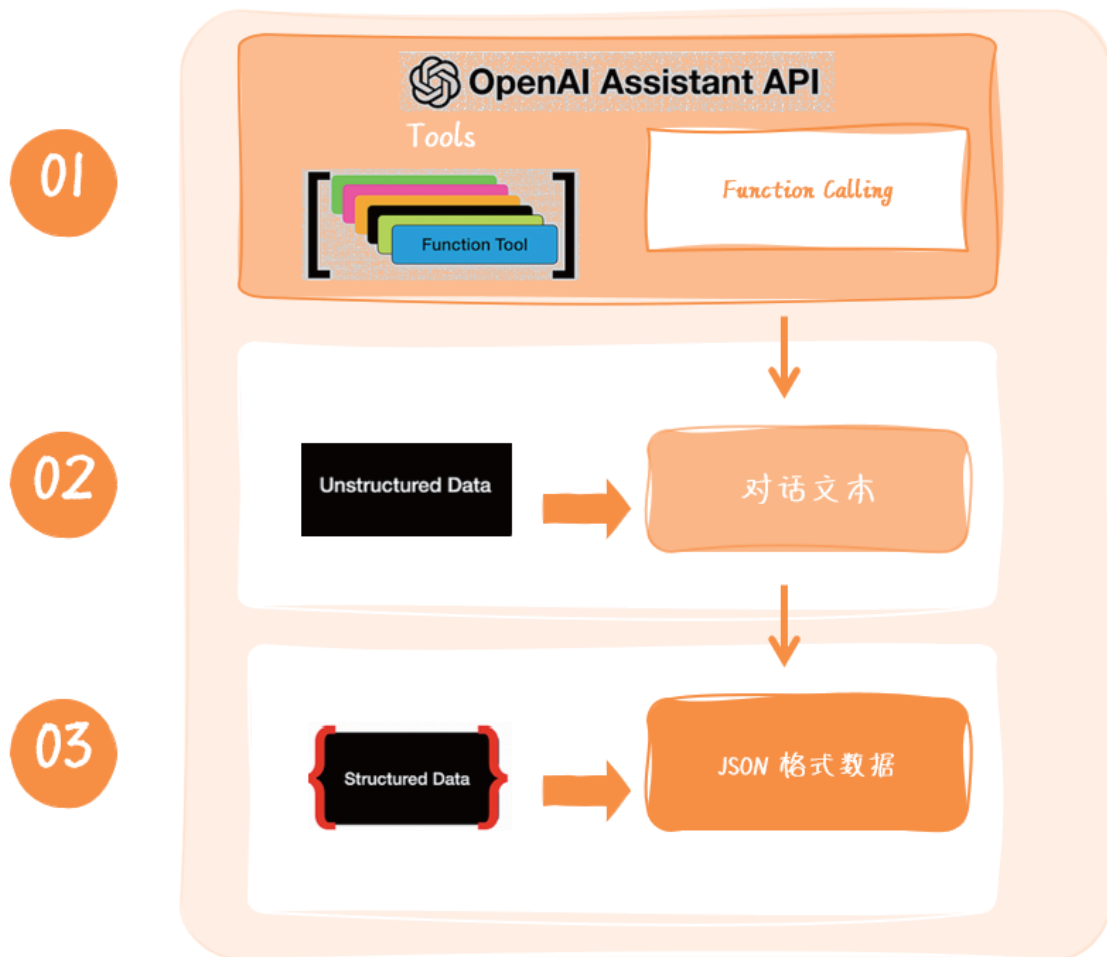


图1.17 会使用工具的Agent

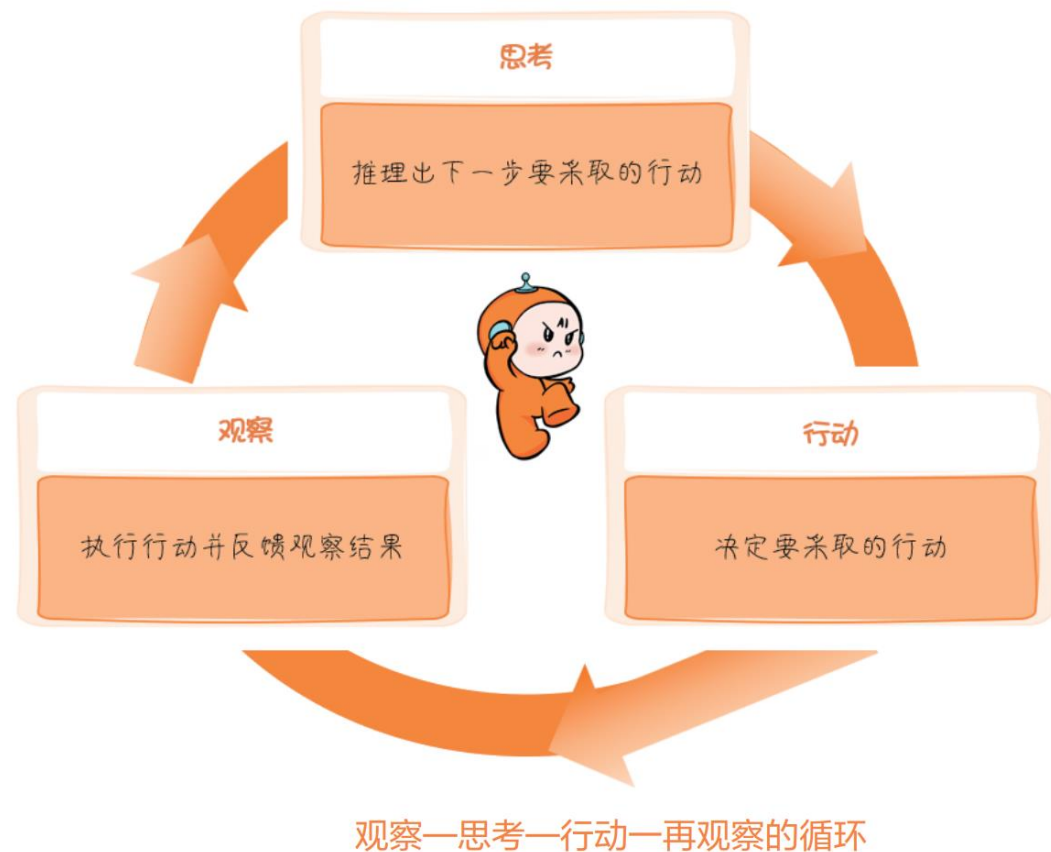
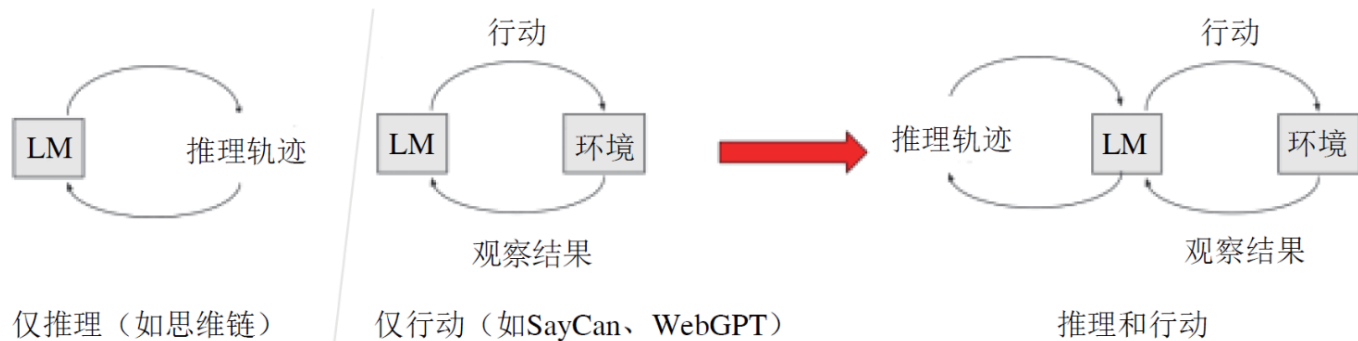


Agent认知框架：Function Callings/Tool Calls



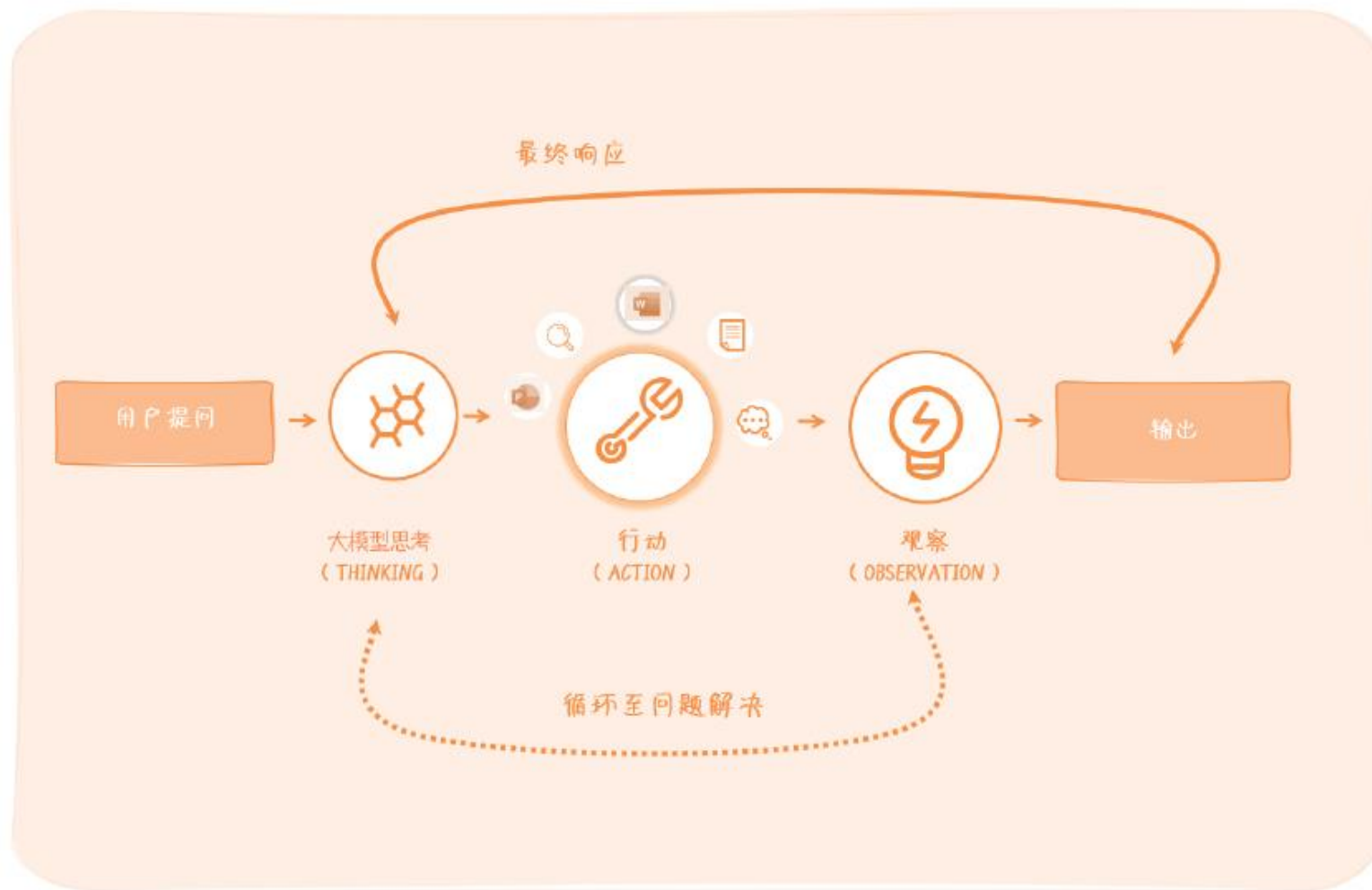
大模型可以通过外部工具获取额外信息

Agent认知框架：ReAct

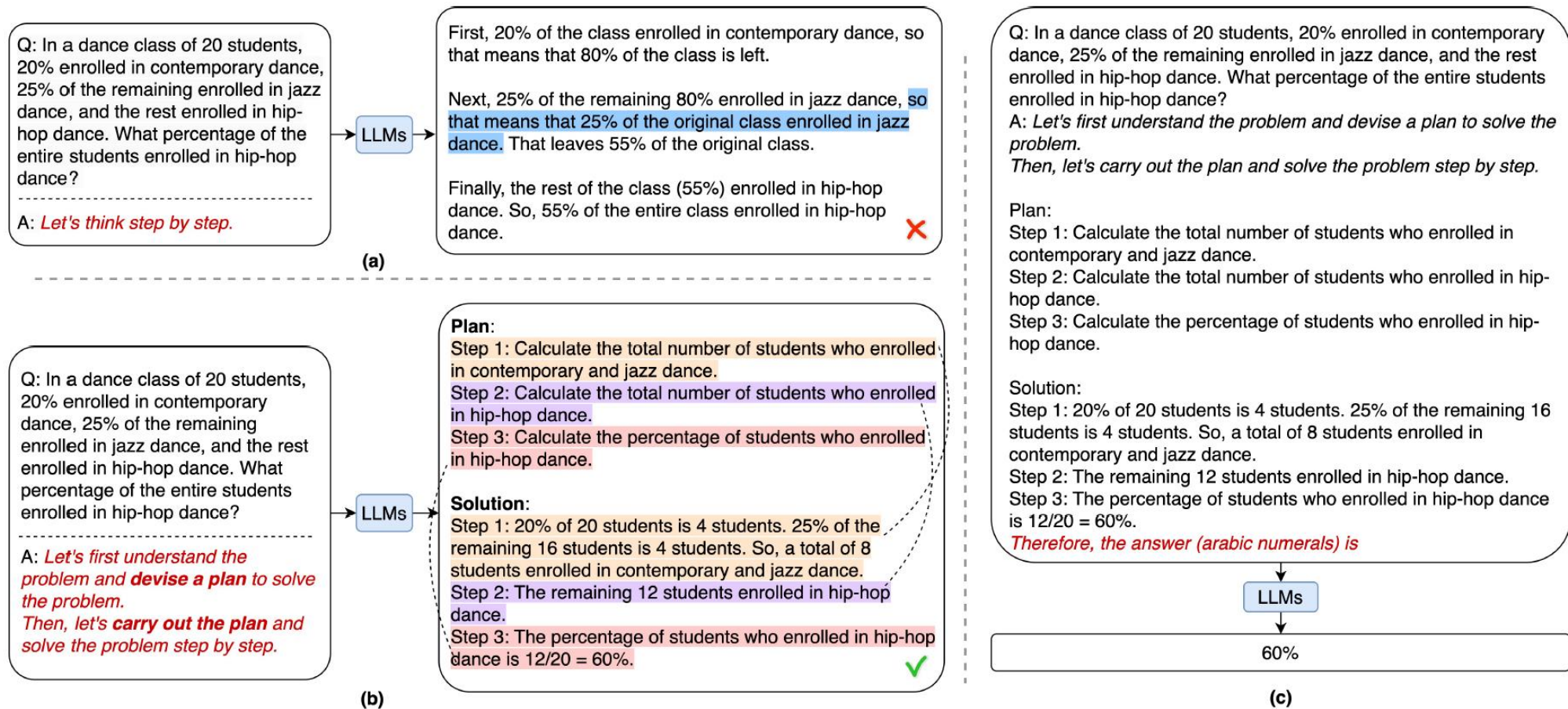


Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*.

▶ Agent认知框架：ReAct



Agent认知框架: Plan-and-Solve



Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., & Lim, E.-P. (2023). Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. *arXiv*.

Agent认知框架：Plan-and-Solve

In

```
# 运行 Agent 解决新问题（完善需求）  
agent.run(" 查查玫瑰花的库存然后给出 50 朵玫瑰花的价格和当天的配送方案！ ")
```

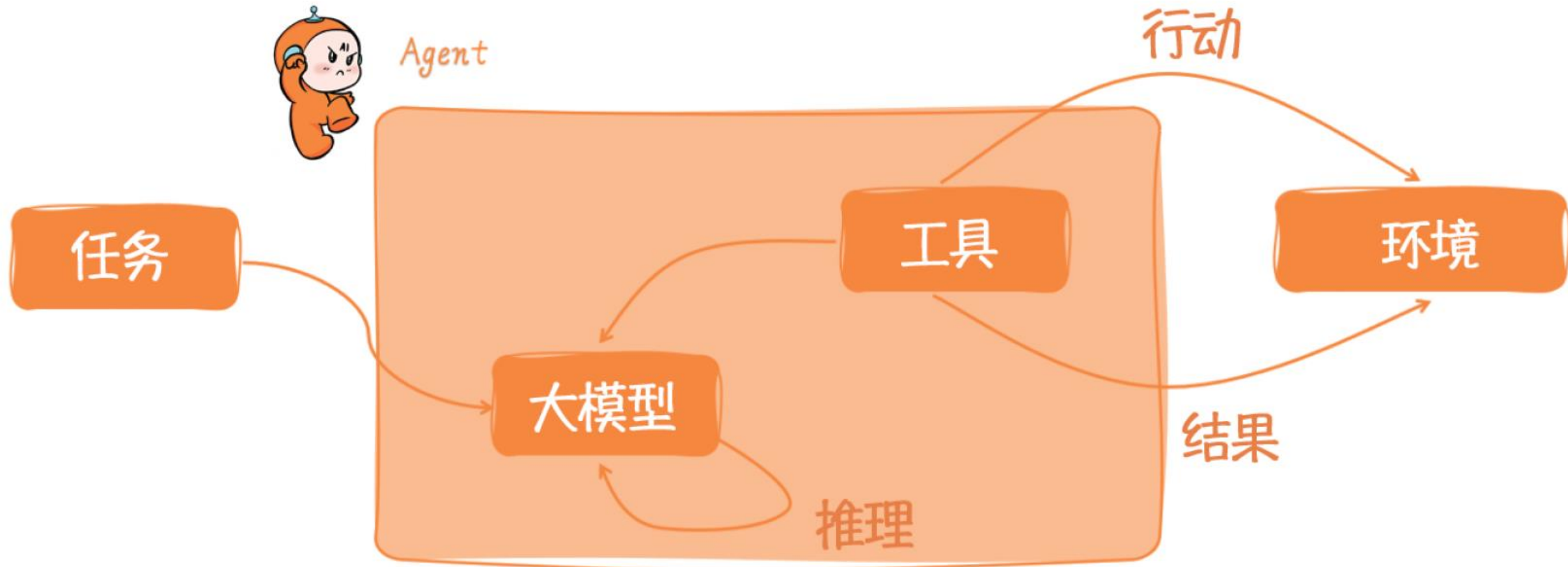
Out

```
> Entering new PlanAndExecute chain...  
steps=[  
Step(value='Check the inventory of roses.'),  
Step(value='If the inventory is sufficient (at least 50 roses), proceed to step 3.  
there are not enough roses in stock and end the conversation.'),  
Step(value='Retrieve the price of 50 roses.'),  
Step(value='Retrieve the delivery options available for the current day.'),  
Step(value='Provide the user with the price of 50 roses and the available delivery
```



Agent在计划阶段给出的新需求的执行步骤

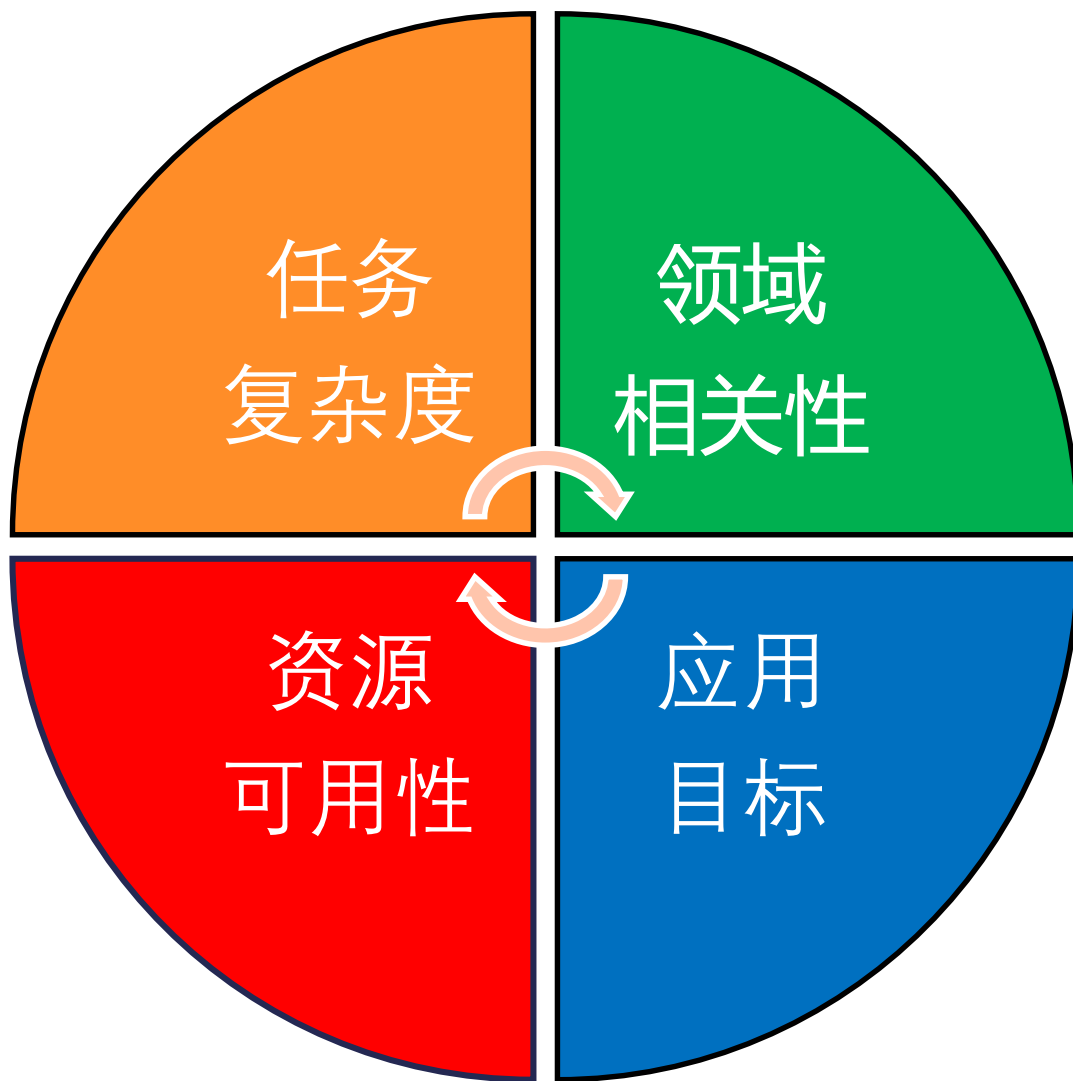
▶ 认知框架的组合：ReAct+Tool Calls



PART 03

Agenti认知框架适用场景分析

▶ 适用场景分析维度



▶▶ 任务复杂度:认知框架的首要选择维度

- 规则是否明确?
- 步骤是否固定?
- 是否涉及分析、推理和决策?

▶▶ 领域相关性:通用还是专用

- 通用领域:如日常对话、文本摘要
- 特定垂直领域:如医疗、法律等

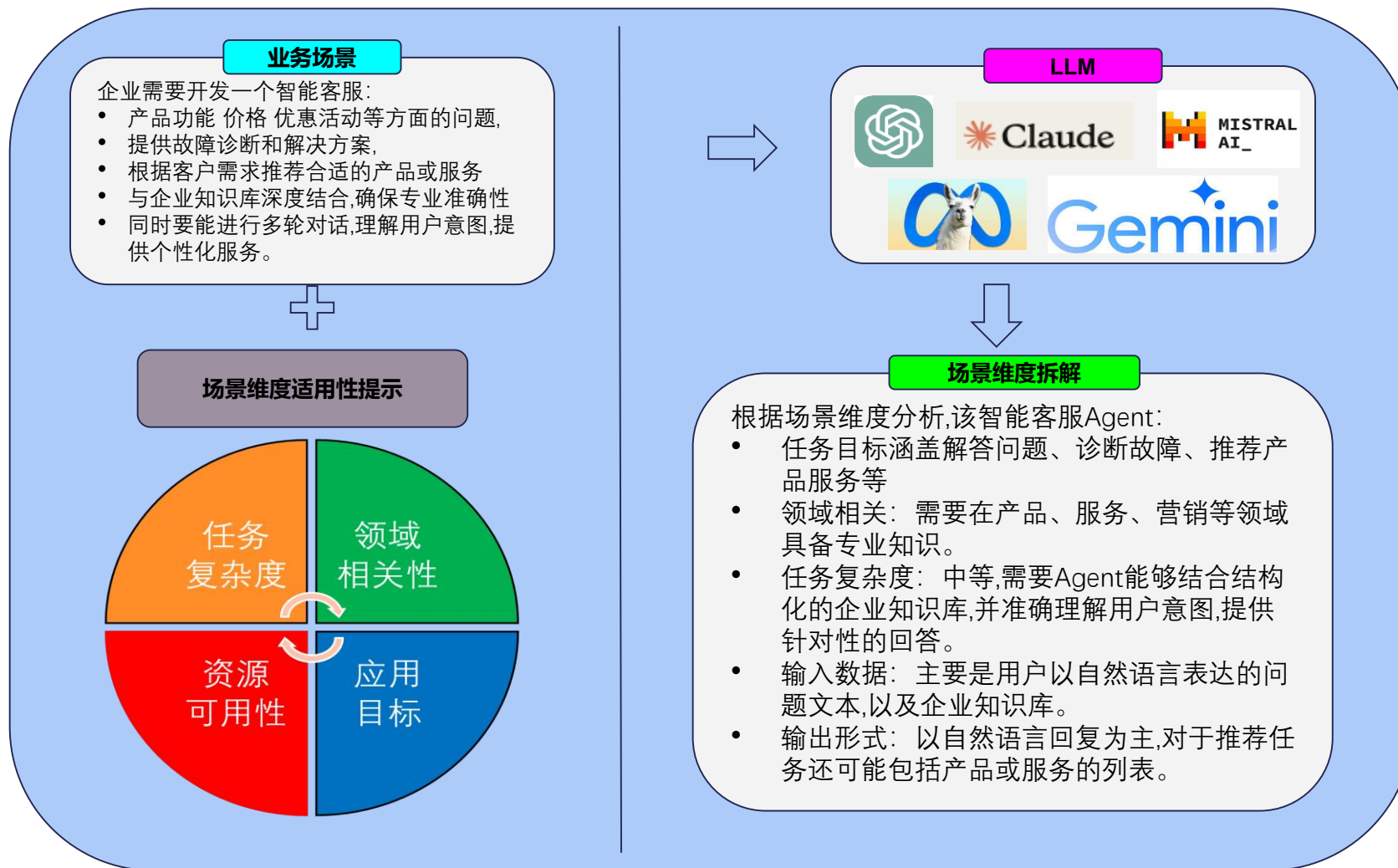
▶▶ 资源可用性:外部信息助力

- 知识库:如百科、图谱
- 语料库:如网页、文档等
- API:如搜索、计算、翻译等

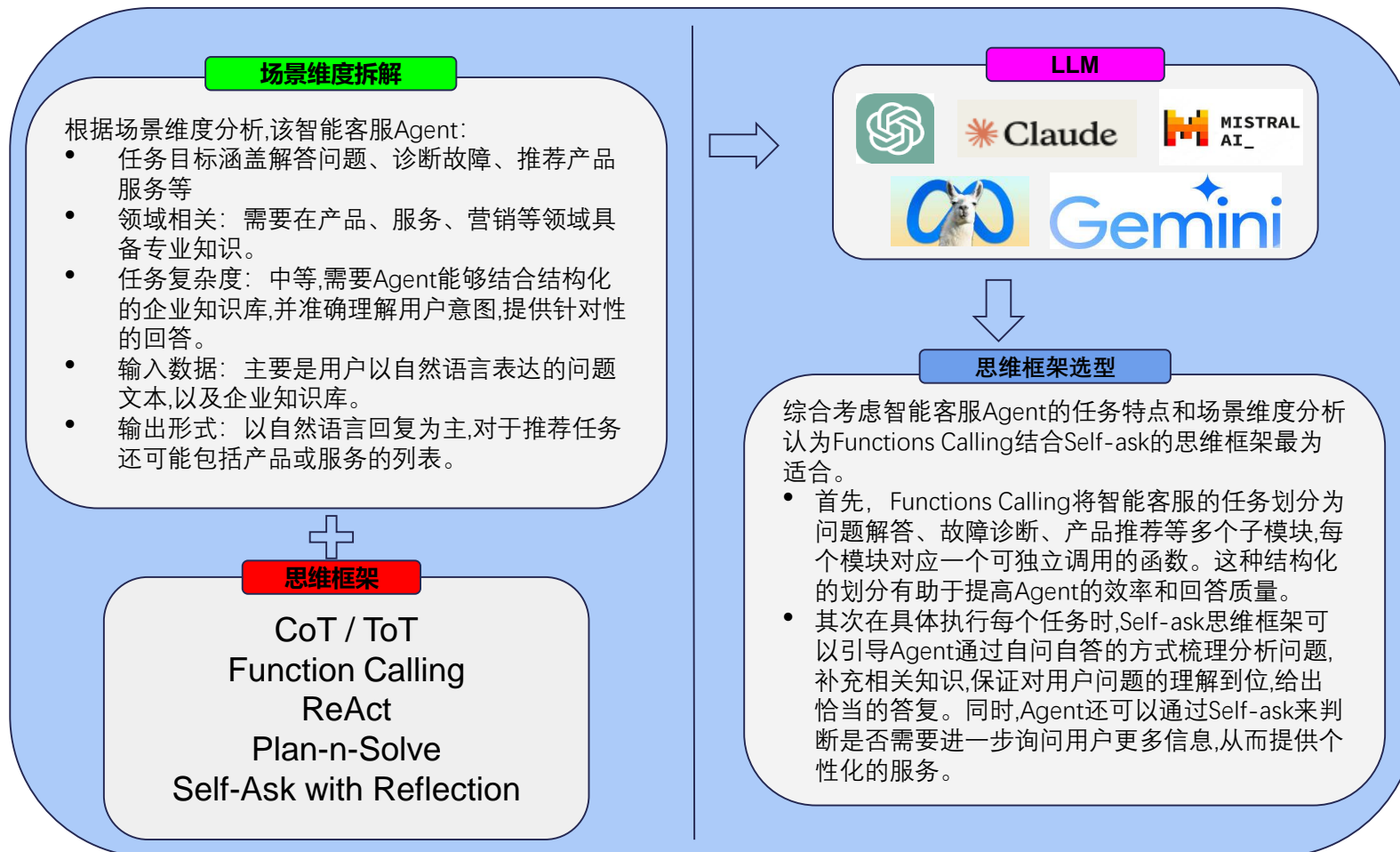
▶▶ 应用目标:根据需求权衡

- 是否追求任务完成的速度和成本?
- 是否追求任务输出的质量和精度?
- 是否追求追求认知过程的透明和可读性?
- 是否追求任务结果的新颖性和发散性?

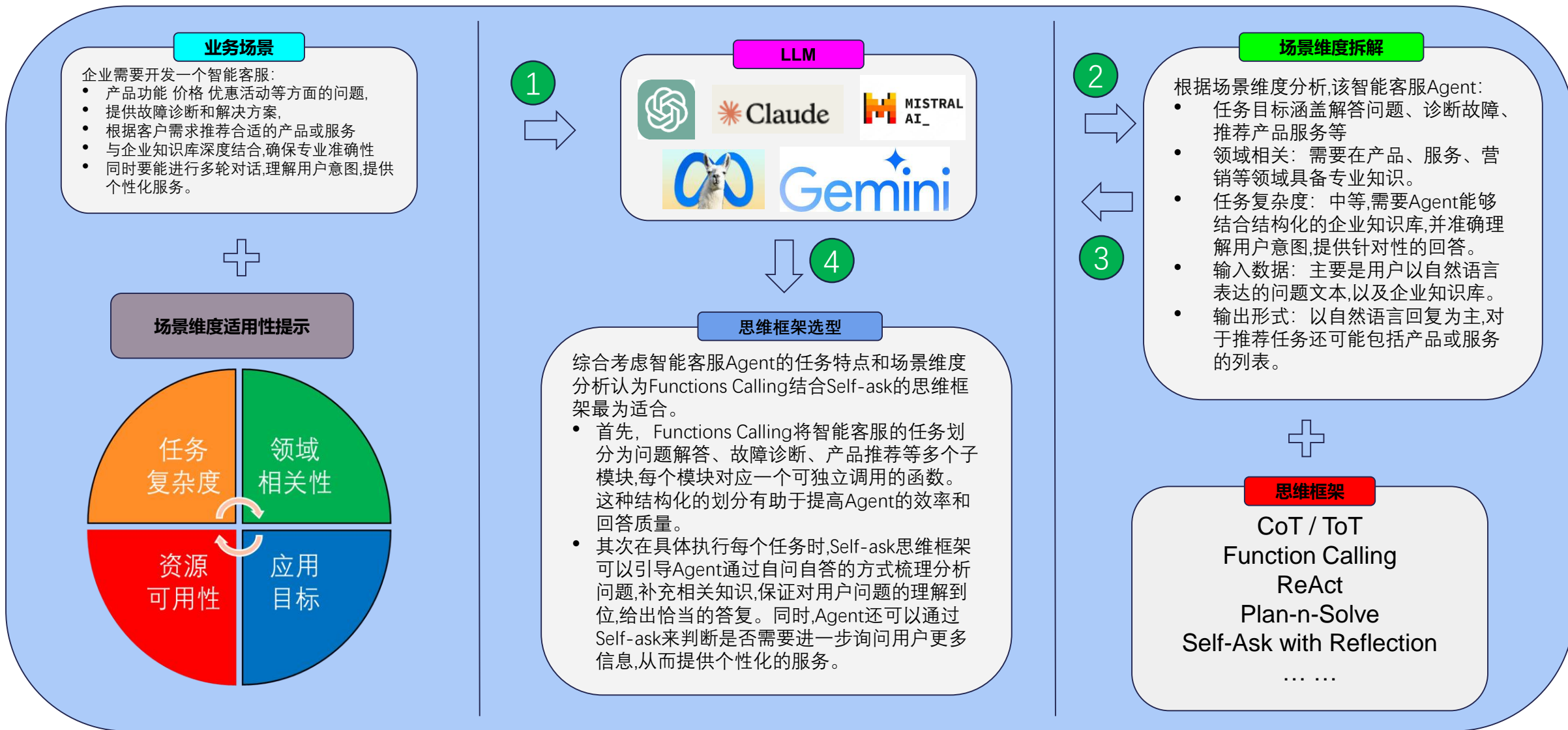
认知框架选型：步骤1 场景维度拆解



认知框架选型：步骤2 思维框架选型



认知框架选型：整体流程



▶ 场景维度提示模板（示意）

```
_Prompt > Thought_Framwork > {} scenario_dimensions.json > ...
```

```
1 {  
2   "Task Objective": "The main goals and objectives of the scenario",  
3   "Task Domain": "The specific domain or industry related to the scenario",  
4   "Task Complexity": "The level of complexity involved in the scenario",  
5   "Input": "The type and format of input data used in the scenario",  
6   "Output": "The expected form and format of the output"  
7 }
```

```
_Prompt > Thought_Framwork > {} 场景维度.json > ...
```

```
1 {  
2   "任务目标": "场景的主要目标和目的",  
3   "任务领域": "与场景相关的特定领域或行业",  
4   "任务复杂度": "场景中涉及的复杂程度",  
5   "输入数据": "场景中使用的输入数据类型和格式",  
6   "输出形式": "期望的输出形式和格式"  
7 }
```

▶ 认知框架提示模板 (示意)

```
Prompt > Thought_Framework > {} cognitive_frameworks.json > ...
```

```
1  {
2      "Functions Calling": "Decompose complex tasks into callable subfunctions",
3      "Chain-of-Thought (CoT)": "Solve complex problems through step-by-step reasoning",
4      "Tree-of-Thought (ToT)": "Explore multiple thought branches in parallel and select the best solution",
5      "ReAct": "Continuously think, interact, and act to complete tasks",
6      "Self-Ask": "Guide thinking and supplement knowledge through self-questioning and answering",
7      "Plan-and-Execute": "Plan first, then execute tasks with continuous monitoring"
8  }
```

```
Prompt > Thought_Framework > {} 认知框架.json > ...
```

```
1  {
2      "功能调用": "将复杂任务分解为可调用的子函数",
3      "思维链(CoT)": "通过逐步推理解决复杂问题",
4      "思维树(ToT)": "并行探索多个思路分支,选择最优解",
5      "反应(ReAct)": "持续思考、交互、行动以完成任务",
6      "自问自答(Self-Ask)": "通过自我提问和回答引导思考并补充知识",
7      "规划执行(Plan-and-Execute)": "先制定计划,再执行任务,持续监控"
8  }
```

▶ 语言模型输出 (示意)

场景维度拆解:

场景维度的详细拆解:

任务目标:

1. ****客户支持****: 智能客服助手应能回答有关公司产品和服务的查询, 确保用户能够获得及时和准确的信息。
2. ****个性化推荐****: 根据用户的查询和反馈, 智能助手应提供个性化的产品或服务建议, 增加用户满意度和可能的销售机会。
3. ****收集反馈****: 助手需要能够收集用户的反馈, 并将这些信息用于未来服务的改进。
4. ****数据分析****: 对收集的数据进行分析, 以便更好地理解客户需求和优化服务。

任务领域:

- ****客户服务****: 主要关注提供高效和有效的客户支持。
- ****数据分析****: 分析用户行为和反馈, 以提供更精准的服务。
- ****人工智能/机器学习****: 利用AI技术进行个性化推荐和自然语言处理。

任务复杂度:

- ****复杂性高****: 智能客服助手需要理解和处理自然语言输入, 生成准确且相关的回答, 进行情感分析以理解用户情感, 并根据复杂的用户数据制作个性化推荐。
- ****需处理的数据类型多样****: 包括文本数据、用户行为数据、购买历史等。
- ****需要高级的算法****: 例如自然语言处理(NLP)、机器学习模型(用于推荐系统和预测分析)。

输入数据:

- ****文本输入****: 用户通过聊天界面输入的自然语言文本。
- ****用户行为数据****: 用户的浏览历史、购买历史和之前的反馈。
- ****客户服务历史记录****: 先前的查询和响应数据, 用于训练AI模型。

输出形式:

- ****文本响应****: 智能助手通过文本回复用户的查询和建议。
- ****数据报告****: 关于用户行为和反馈分析报告, 供公司内部使用, 以改进产品和服务。
- ****推荐列表****: 针对用户可能感兴趣的产品或服务的建议清单。

整体来看, 此智能客服系统将依靠先进的技术和复杂的数据处理能力, 以提供高质量的用户交互体验和增强的业务洞察力。

▶ 语言模型输出 (示意)

选择的认知框架：

在这种复杂的智能客服助手场景下，最适合选择的认知框架是**思维树(ToT)**。思维树可以帮助你并行探索多个思路分支，并在其中选择最优解。在处理涉及客户支持、个性化推荐、数据分析等多方面的复杂任务时，需要同时考虑不同的因素和可能的解决方案。


通过思维树的方法，你可以同时思考如何提高客户支持的效率、采用何种算法来实现个性化推荐、如何分析用户数据以优化服务等多个方面。随着对不同分支的思考，你可以逐步剔除不太适合的方案，最终选择最适合当前情况的最佳解决方案。

这种方法还能帮助你在处理复杂任务时更全面地考虑各种可能性，确保不会遗漏重要的细节，同时也有助于提高决策的准确性和效率。通过思维树，你可以更好地管理和应对这种涉及多个方面的复杂智能客服任务。

PART 04

基于主流框架的Agent实现案例

常用的Agent应用开发框架

 LangChain



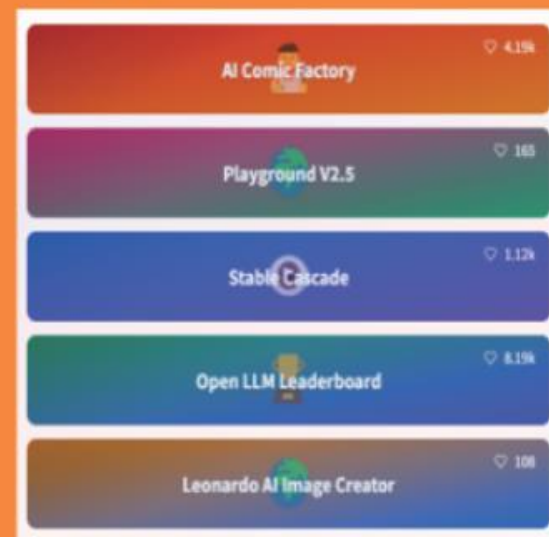
 LlamaIndex



OpenAI API



 Hugging Face



▶ 基于OpenAI Assistants的自动办公助理



项目背景:

客户是采购了GPT-4的企业用户, 为了提升办公效率, 降低人力成本, 客户邀请我们共同开发自动数据分析和办公自动化项目。

- **自动化:** 利用OpenAI的自然语言处理和生成能力, 自动完成文档撰写、数据分析、PPT创作等任务。
- **减少人工:** 通过Assistants API与企业内部系统对接, 实现工作流程的自动化, 大幅减少人工操作环节。
- **复杂场景:** 在收集销售数据的基础上, 进行数据分析, 计算每个季度的销售额, 生成销售数据趋势图表, 并创建PPT。



解决方案:

框架选择:

通过不断比对不同框架效果, 考虑到任务复杂度、领域相关性、资源可用性和项目目标, 我们最终选择基于OpenAI Assistants开发框架通过Tool Calls(Code Interpreter, Function Calling等工具)来建立Agent。

开发框架

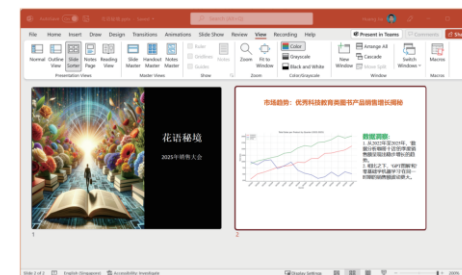
认知框架

业务流程

数据准备

项目实施

您的 Assistant 正在努力做图表呢 ...
您的 Assistant 正在努力做图表呢 ...
您的 Assistant 正在努力做图表呢 ...
您的 Assistant 正在努力做图表呢 ...
.....
图表已创建!



基于OpenAI Assistants的自动办公助理

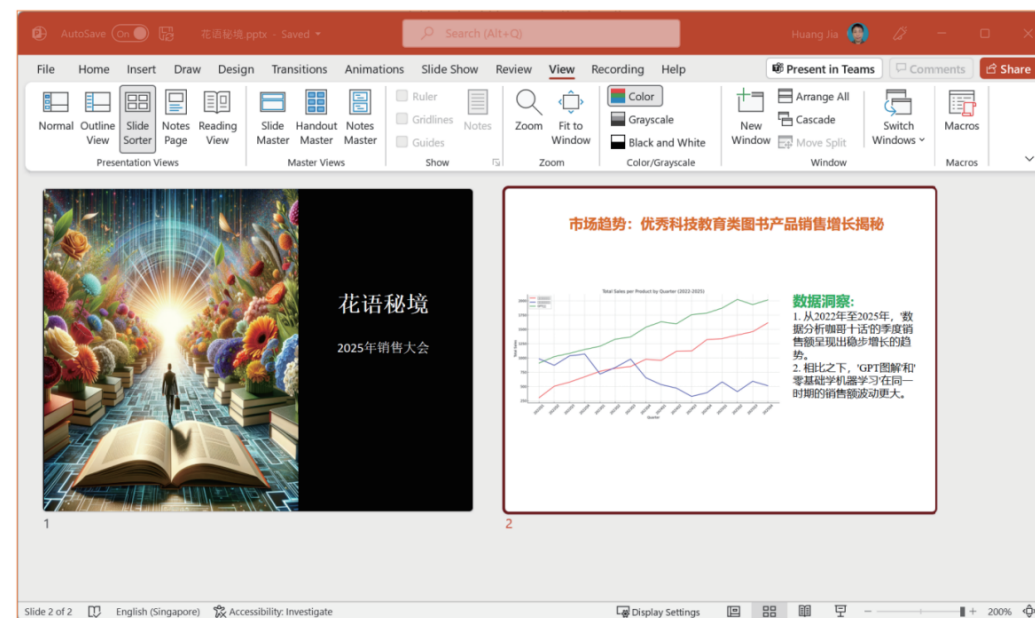
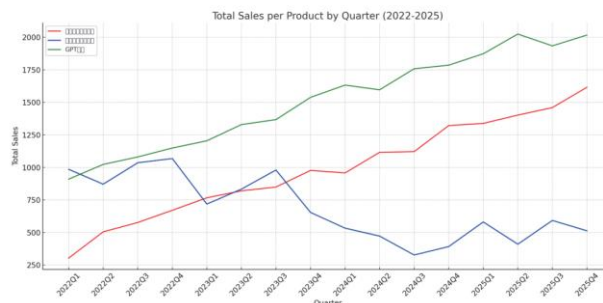
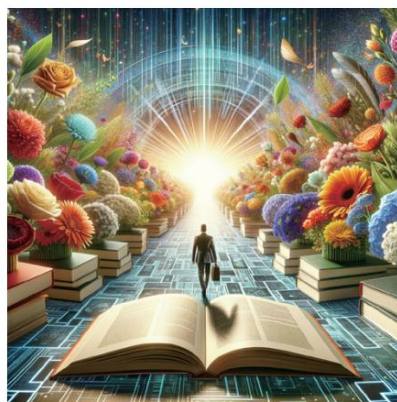


您的 Assistant 正在努力做图表呢 ...
 您的 Assistant 正在努力做图表呢 ...
 您的 Assistant 正在努力做图表呢 ...
 您的 Assistant 正在努力做图表呢 ...

 图表已创建!

“市场趋势：优秀科技教育类图书产品销售增长揭秘” <- (第一次跑的结果)
 “产品销售趋势：稳健增长与市场波动” <- (第二次跑的结果)

日期	零基础学机器学习	数据分析咖哥十话	GPT图解
31/3/2022	303.0327596	985.6150332	909.7627008
30/6/2022	504.8929768	871.1291092	1023.037873
30/9/2022	576.7235126	1035.95079	1080.552675
31/12/2022	670.5018222	1068.18203	1148.976637
31/3/2023	766.7927513	718.3524294	1204.73096
30/6/2023	819.8737846	833.1219711	1329.178823
30/9/2023	849.2219043	980.0890936	1367.517442
31/12/2023	977.0793764	654.1820299	1538.3546
31/3/2024	957.7411639	533.274624	1632.732552
30/6/2024	1115.988153	472.274624	1596.688304
30/9/2024	1121.502993	326.9059767	1758.345008
31/12/2024	1321.700338	391.7368329	1785.778984
31/3/2025	1338.277248	580.9623858	1873.608912
30/6/2025	1402.199291	409.9209298	2025.119328
30/9/2025	1459.683342	592.4668275	1934.207212
31/12/2025	1616.135053	512.8926298	2017.42586



基于LangChain的ReAct Agent实现



项目背景:

某在线鲜花平台，希望开发智能定价系统，让 Agent 可以根据实时的天气和交通状况自动调整产品价格，从而优化销售策略和库存管理：

- **大模型**：利用大语言模型的能力,进行推理，自动判断需要调用哪些工具来完成的任务。
- **工具调用**：利用Google搜索等工具查询天气，交通等信息，作为自动定价系统的参考。
- **ReAct**：通过推理-行动-观察的循环，反复确认任务是否已经成功完成。



解决方案:

框架选择:

通过不断比对不同框架效果，考虑到任务复杂度、领域相关性、资源可用性和项目目标，我们最终选择基于LangChain开发框架通过ReAct思维模式来建立Agent。

开发框架

认知框架

业务流程

数据准备

项目实施

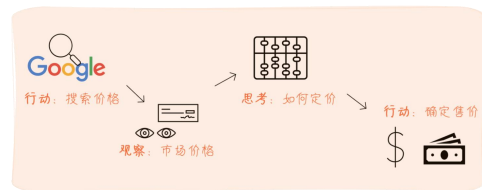


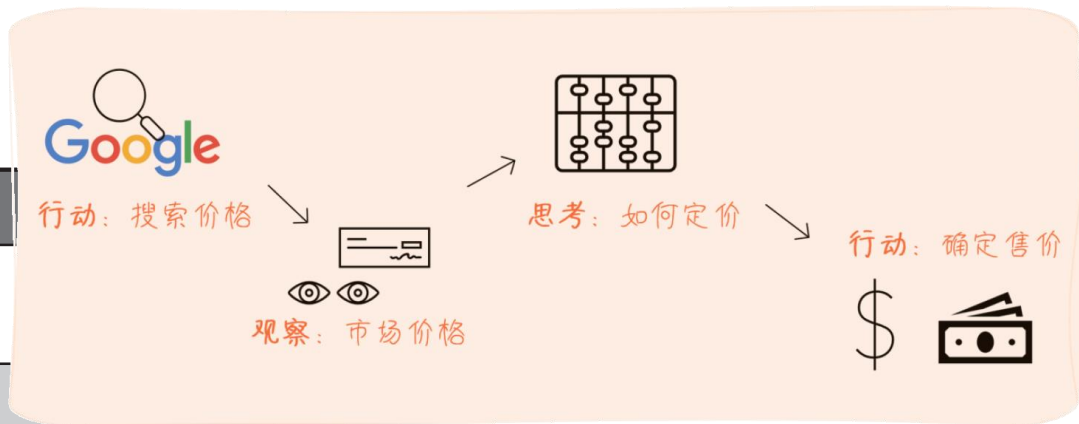
表 6.5 Agent 的思考-行动链条

步骤编号	中文说明	步骤内容	详细描述
1	开始	Entering new AgentExecutor chain...	开始一个新的智能 Agent 执行链
2	行动	Action: Search	智能 Agent 准备执行搜索操作
3	行动输入	Action Input	输入了搜索指令：“玫瑰市场平均价格”
4	观察	Observation	获取了有关玫瑰花市场价格的详细信息，包括价格波动的原因和零售价格范围
5	思考	Thought	Agent 反思需要计算玫瑰价格加价 5% 后的数额
6	行动	Action: Calculator	决定使用计算器工具来计算价格加价后的结果
7	行动输入	Action Input	输入计算指令“51 * 1.05”，用于计算价格加价 5%
8	观察	Observation	观察到计算结果是 5.25，表明玫瑰花价格加价 5% 后的价格为 55.25
9	最终答案	Final Answer	确定并提供最终答案：玫瑰花价格加价 5% 后为 55.25
10	结束	Finished chain	表示智能 Agent 的任务链已经完成

▶ 基于LangChain的ReAct Agent实现

表 6.5 Agent 的思考 – 行动链条

步骤编号	中文说明	步骤内容	详细描述
1	开始	Entering new AgentExecutor chain...	开启一个新的智能 Agent 执行链
2	行动	Action: Search	智能 Agent 准备执行搜索操作
3	行动输入	Action Input	输入了搜索指令：“玫瑰市场平均价格”
4	观察	Observation	获取了有关玫瑰花市场价格的详细信息，包括价格波动的原因和零售价格范围
5	思考	Thought	Agent 反思需要计算玫瑰价格加价 5% 后的数额
6	行动	Action: Calculator	决定使用计算器工具来计算价格加价后的结果
7	行动输入	Action Input	输入计算指令“ $51 * 1.05$ ”，用于计算价格加价 5%
8	观察	Observation	观察到计算结果是 5.25，表明玫瑰花价格加价 5% 后的价格为 55.25
9	最终答案	Final Answer	确定并提供最终答案：玫瑰花价格加价 5% 后为 55.25
10	结束	Finished chain	表示智能 Agent 的任务链已经完成



▶ 基于LangChain的ReAct Agent实现

> Entering new AgentExecutor chain...

我需要先找到目前市场上玫瑰花的一般进货价格，然后在这个价格基础上加价5%来计算最终的定价。

Action: Search

Action Input: 目前市场上玫瑰花的一般进货价格['记者昨日在上海双季花卉批发市场看到，众多鲜切花批发商家都在为即将到来的“双节”做准备。几天前叫价2元一支的红色、粉色玫瑰花，目前的零售价已渐行上涨。在一些街道的花店，同类玫瑰花的价格已卖到3元一支，更高的甚至还要4元。', '批发市场价格上涨，鲜花零售门店价格更为高昂。位于朝阳大悦城的一家花店工作人员向北京商报记者介绍道，“之前一枝玫瑰(不分品种)的价格是25元，现在已经 ...', '除了产自昆明的玫瑰外，广州、佛山的本地玫瑰价格也有所增长，但在玫瑰普遍涨价的趋势下，本地玫瑰的价格仍然具有优势。经营鲜切花店十多年的王女士称，广州本地产的一扎(20枝)的C级红玫瑰，平日的价格是20至30元，今年的普遍批发价格是120元，最贵的已叫价150元。', '今年的玫瑰价格比往年几乎翻了一番，原因是什么？不少花店老板都表示，他们的进货地是云南昆明，今年因为天气原因，昆明的玫瑰产量跟不上。“今年昆明冻了 ...', '214即将到来，鲜花销售迎来旺季。但今年最能代表爱情的玫瑰花价格暴涨四五倍，让不少商家都高兴不起来。不少商户表示，进货时都是颤抖着双手付款的。', '... 玫瑰的单支价格约为10-12元，今年红玫瑰单支价格普遍在20元以上，新晋“网红”卡布奇诺玫瑰，去年单支价格约在15-20元，今年基本上进货价都在25-30元。', '今年由于气温回升较快，玫瑰花已经开始采摘。平阳县玫瑰鲜花蕾开秤价格从2.8元/斤，迅速攀升至4月27日的3.5元/斤，开秤价格平均每斤比去年同期上涨1元，涨幅约为66.67% ...', '... 花要卖980元。玫瑰价格疯涨。上海情人节的“玫瑰仪式感”依旧丝毫未减。春寒料峭，花店、批发市场门庭若市，记者向多家摊主老板了解到，“今年的玫瑰花 ...', '普通玫瑰花一般在市场上售价为3~5元一朵，其他玫瑰的价格有所变化，越珍贵的玫瑰品种价格越高，比如蓝色妖姬在市场上售价就能达到几十元一朵。生活中的 ...', '... 价大约在每支2-3元，现在已经涨到了6元。杨先生从2015年开始经营鲜花批发生意，一般提前15天为情人节备货，由于鲜花价格上涨，今年备货压力陡增。“批发 ...']从观察到的信息中，可以看到玫瑰花的进货价格有很大的波动，具体价格取决于品种、地区和时间。为了简化计算，我们可以取一个中间值作为进货价格的参考。根据观察到的信息，普通玫瑰花的进货价格大约在25-30元之间。我们可以取这个范围的中间值，即27.5元作为计算的基础。

Thought: 现在我有了一个大概的进货价格，接下来我需要计算在此基础上加价5%的定价。

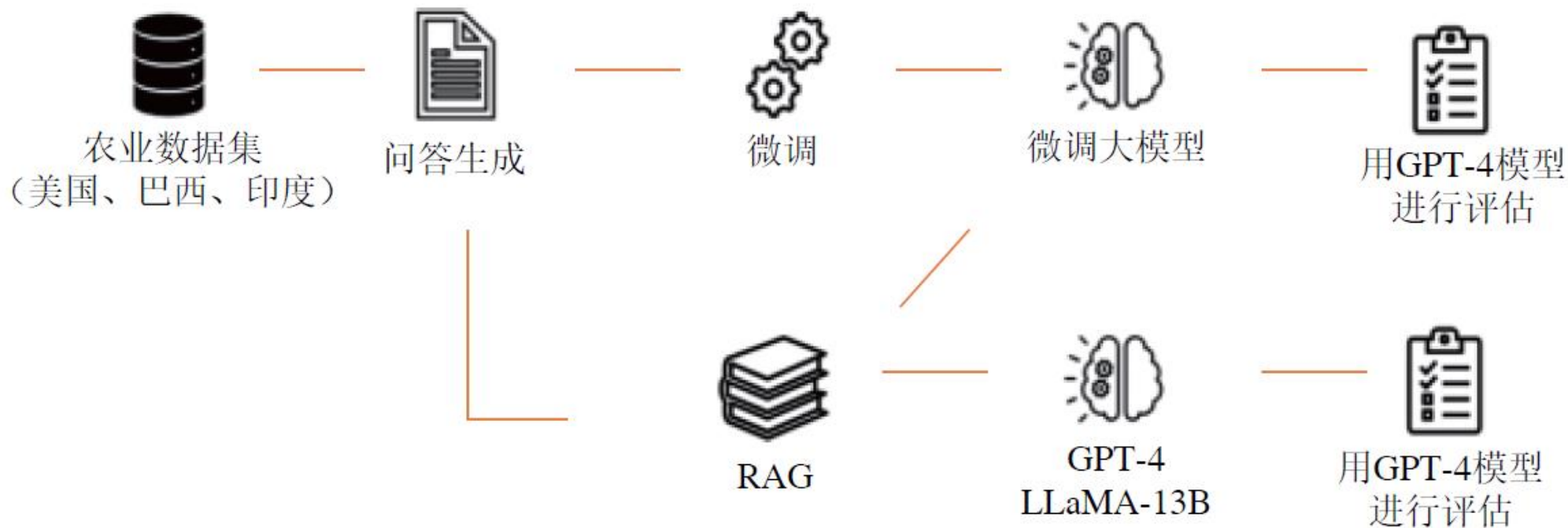
Action: Calculator

Action Input: 27.5 * 1.05Answer: 28.875I now know the final answer

Final Answer: 在进货价格大约为27.5元的基础上加价5%，最终的定价应该是28.88元。

> Finished chain.

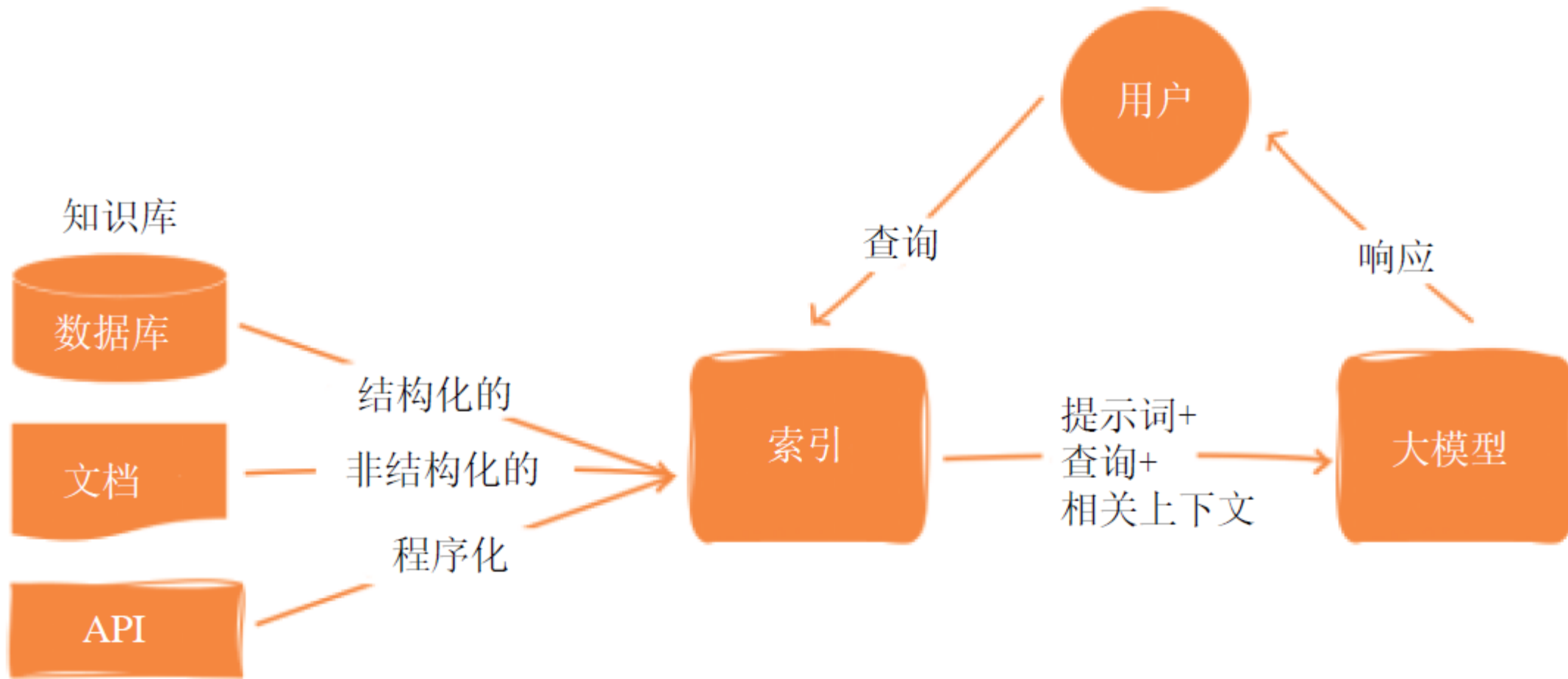
▶ 基于LlamaIndex的RAG ReAct Agent实现



基于农业数据集大模型应用RAG和微调的过程

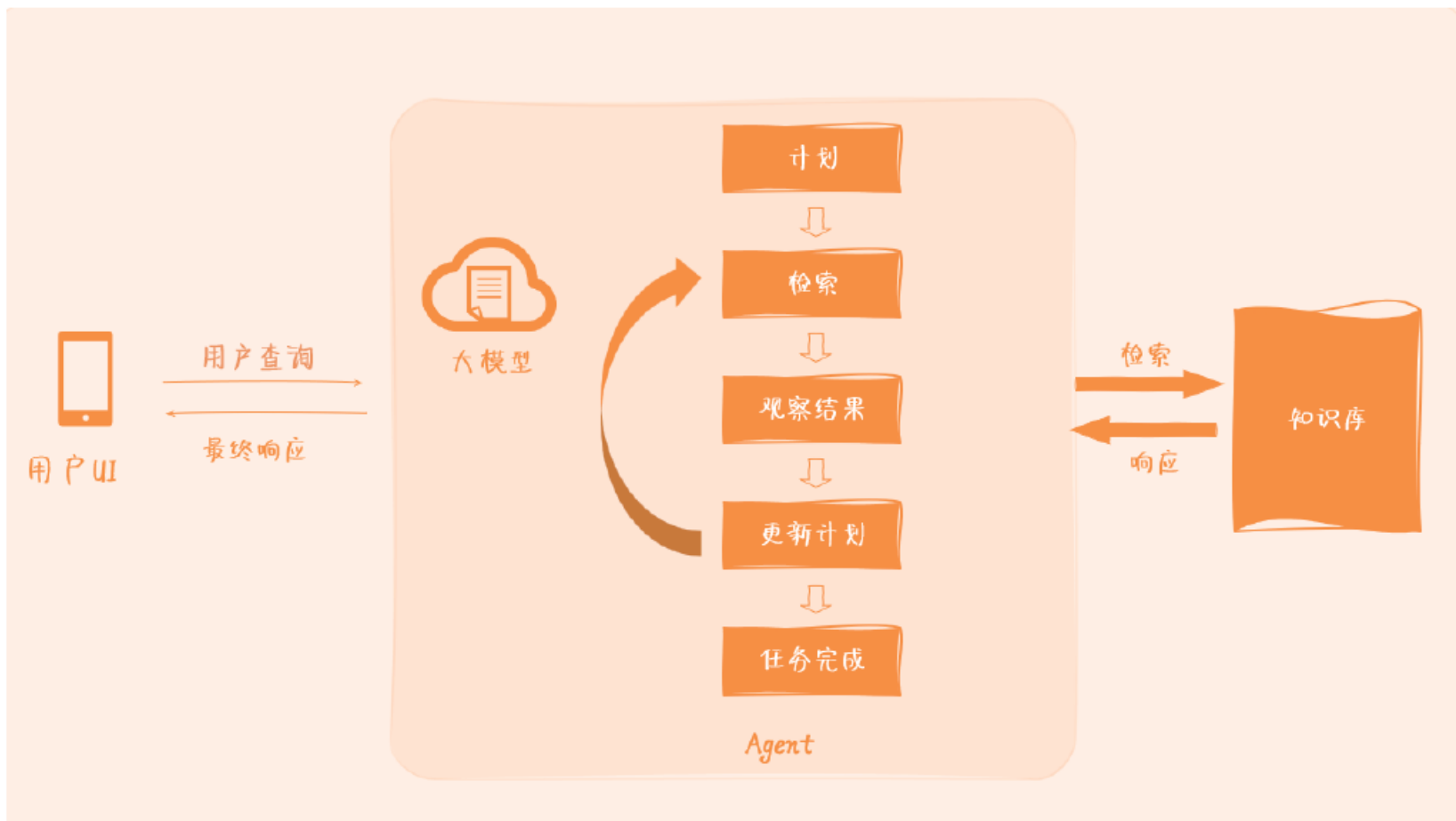
Balaguer, A., Benara, V., & Chandra, R. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. *arXiv*. <https://doi.org/10.48550/arXiv.2401.08406>

▶ 基于LlamaIndex的RAG ReAct Agent实现



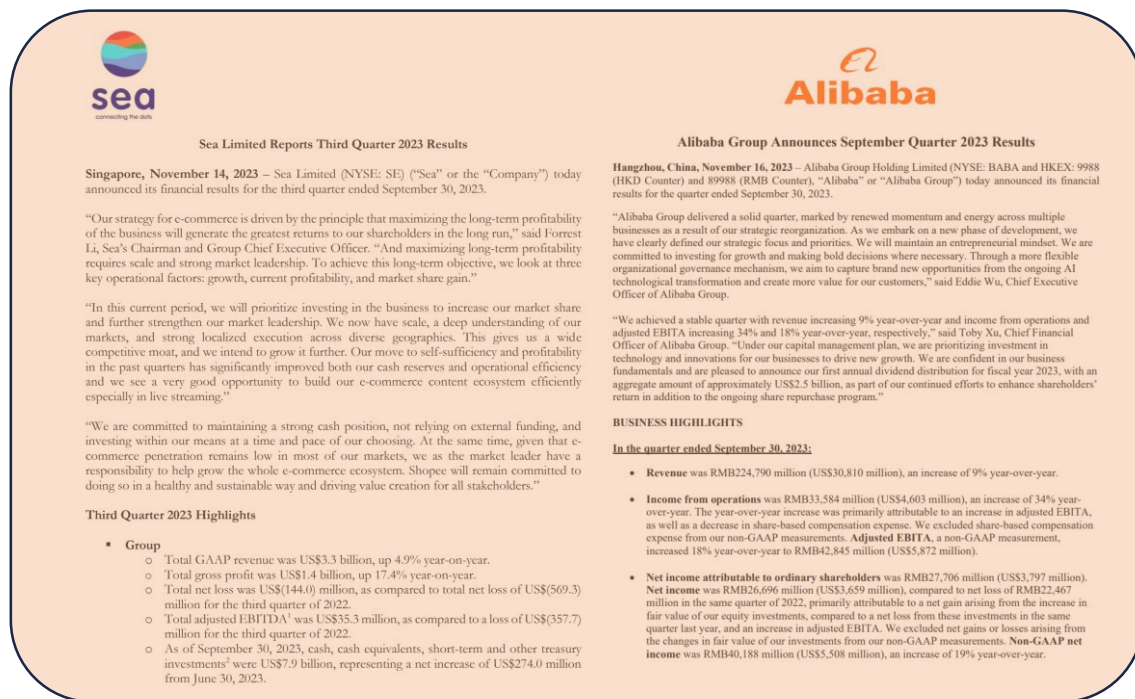
用户视角的RAG流程

▶ 基于LlamaIndex的RAG ReAct Agent实现



融合RAG能力的Agent

▶ 基于LlamaIndex的RAG ReAct Agent实现



Thought: I need to use a tool to help me compare the sales revenue of the two companies.

Action: A_Finance

Action Input: {'input': 'Please provide the sales revenue for both companies.'}

Observation: The sales revenue for the company in the third quarter of 2023 was \$3.3 billion.

Thought: I have obtained the sales revenue for one company, but I still need the sales revenue for the other company to make a comparison.

Action: B_Finance

Action Input: {'input': 'Please provide the sales revenue for the other company.'}

Observation: The sales revenue for the other company in the quarter ended September 30, 2023 was RMB48,052 million (US\$6,586 million).

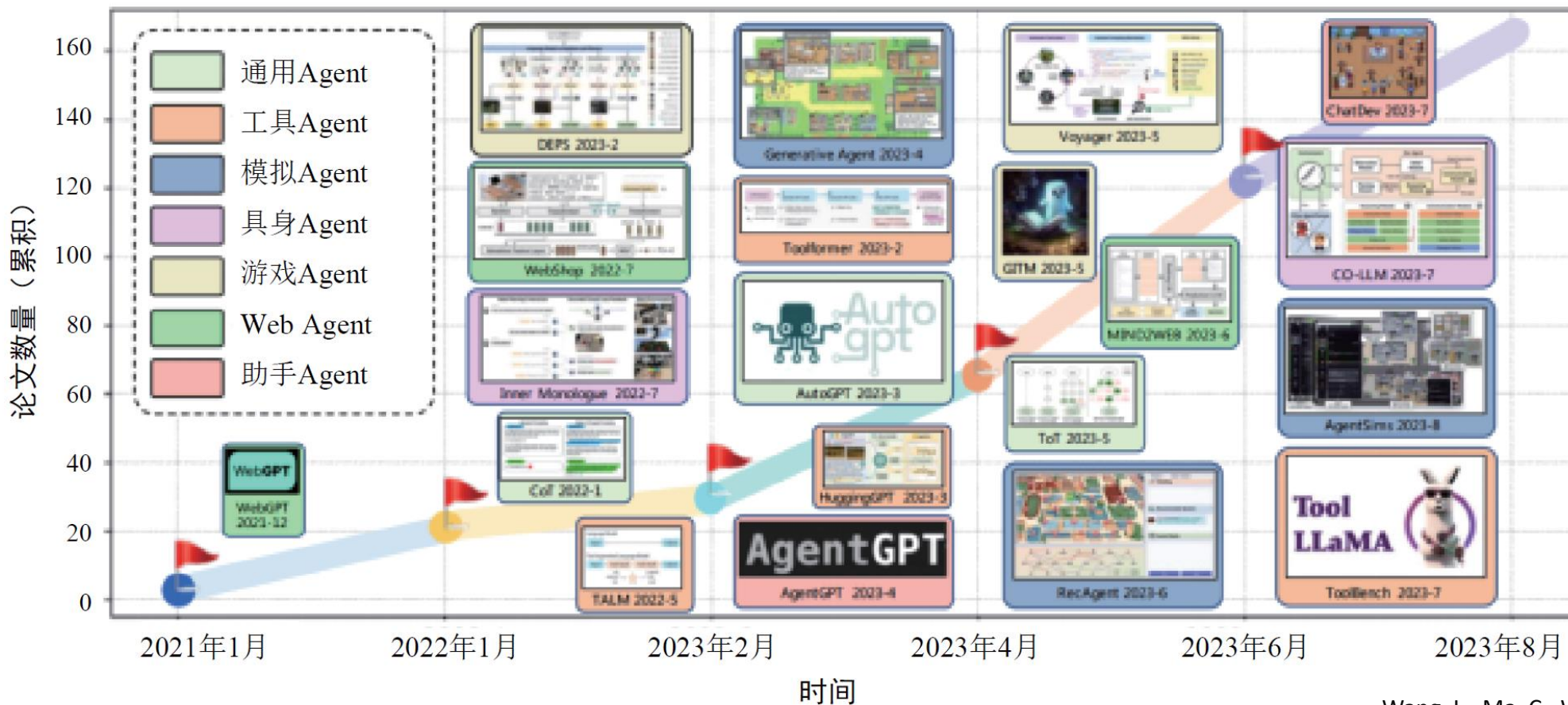
Thought: I have obtained the sales revenue for both companies. Now I can compare them.

Answer: The sales revenue for Company A in the third quarter of 2023 was \$3.3 billion, while the sales revenue for Company B in the same period was \$6.586 billion. Company B had a higher sales revenue compared to Company A.

PART 05

AI Agent总结与展望

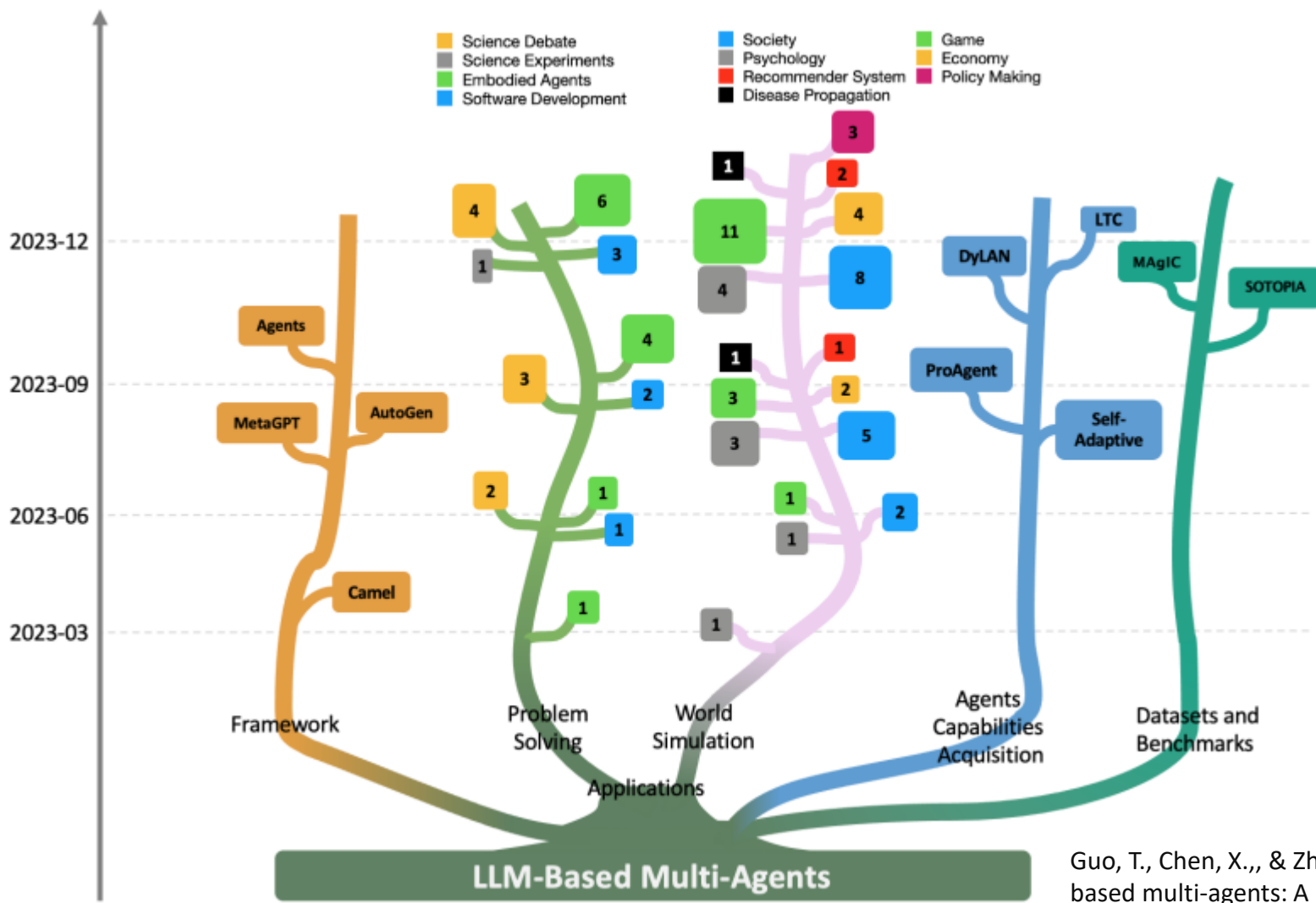
Agent研究综述



从2021年到2023年由大模型驱动的Agent研究的发展脉络

Wang, L., Ma, C., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18.

多Agents研究综述



Guo, T., Chen, X., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges.

▶▶ Agent未来的热点方向?

- Agent 自主学习
- 多 Agent 合作
- Agent 的可信度和安全性评估
- 大模型和 Agent 在边缘系统中的部署
- 具身智能的落地

未来的Agent架构应该具有自我复制和进化能力

博学之，审问之，慎思之，明辨之，笃行之。

——《礼记·中庸》

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **K+峰会**  **上海站**
K+ 全球软件研发行业创新峰会
时间: 2024.06.21-22

 **K+峰会**  **敦煌站**
K+ 思考周®研习社
时间: 2024.10.17-19

 **K+峰会**  **香港站**
K+ 思考周®研习社
时间: 2024.11.10-12



K+峰会详情



 **AiDD峰会**  **上海站**
AI+研发数字峰会
时间: 2024.05.17-18

 **AiDD峰会**  **北京站**
AI+研发数字峰会
时间: 2024.08.16-17

 **AiDD峰会**  **深圳站**
AI+研发数字峰会
时间: 2024.11.08-09



AiDD峰会详情

THANKS

