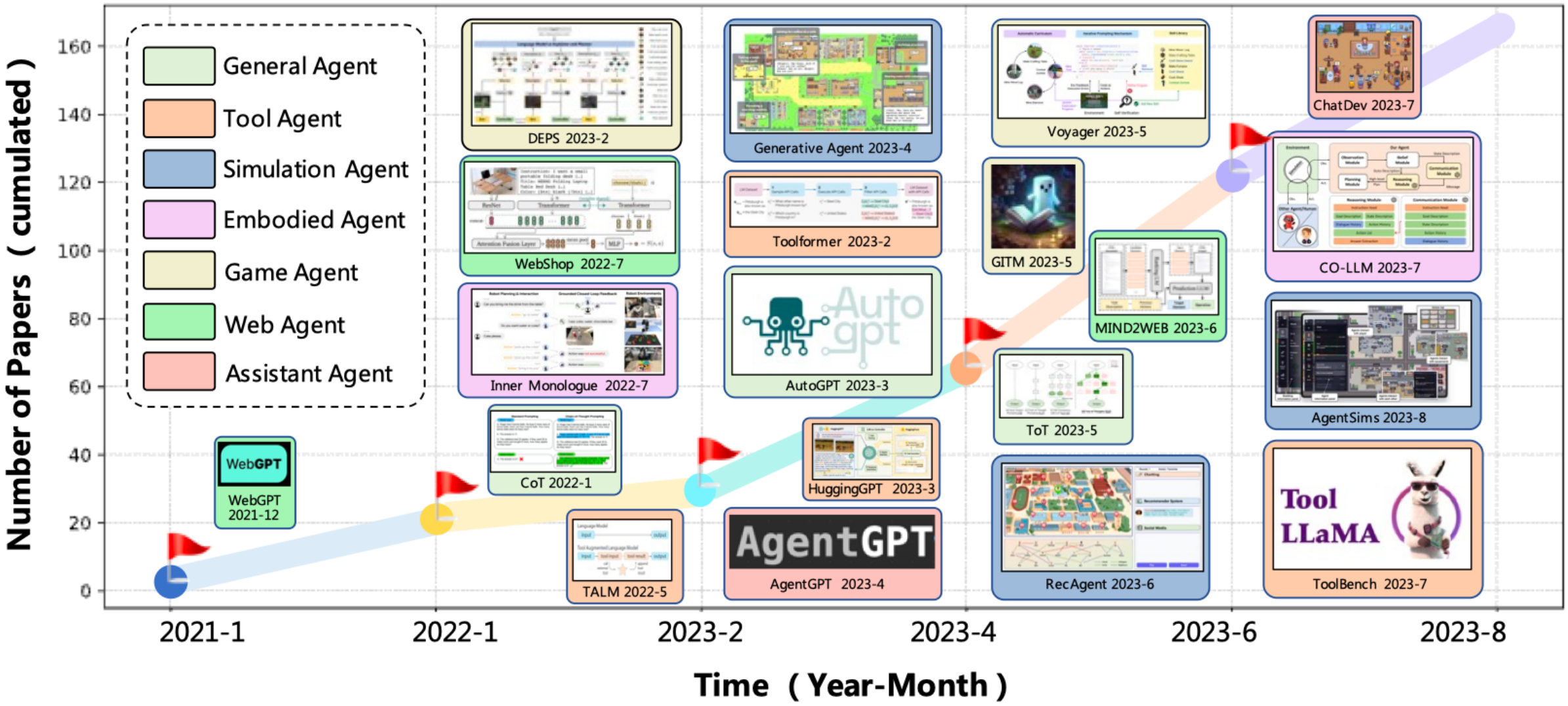

ZOMI
大模型系列之智能体
AI Agent 组成介绍

关于大模型系列

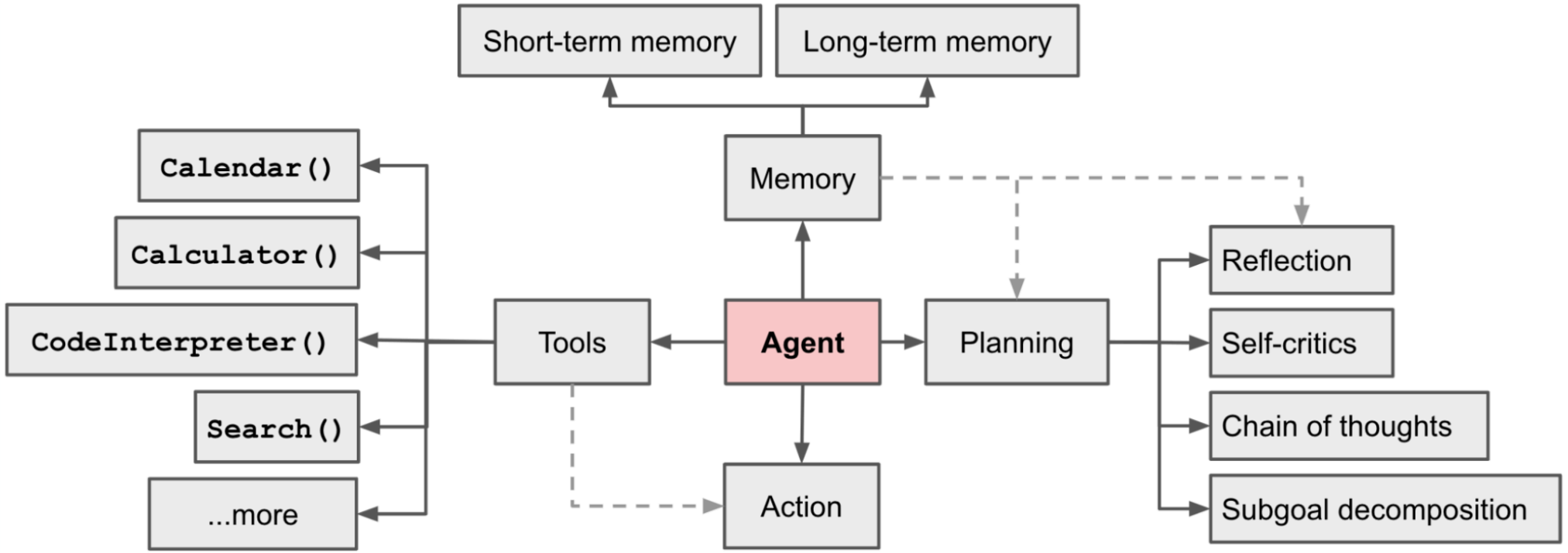
- 内容背景
  - LLM + AI Agent：大模型遇到智能体
- 具体内容
  1. AI Agent 组成介绍：LLM + 记忆 + 规划 + 工具
  2. AI Agent 规划手段： Task Decomposition 与 Self Reflection
  3. AI Agent 热门应用：交互式 Agent、自动化 Agent 与多模态 Agent
  4. AI Agent 问题与挑战： Agent 的问题、Agent 的局限性


General Agent
Tool Agent
Simulation Agent
Embodied Agent
Game Agent
Web Agent
Assistant Agent
Number of Papers (cumulated)
0
20
40
60
80
100
120
140
160
2021-1
2022-1
2023-2
2023-4
2023-6
2023-8
Time (Year-Month)
WebGPT
WebGPT 2021-12
DEPS 2023-2
WebShop 2022-7
Inner Monologue 2022-7
CoT 2022-1
TALM 2022-5
Generative Agent 2023-4
Toolformer 2023-2
AutoGPT 2023-3
HuggingGPT 2023-3
AgentGPT
AgentGPT 2023-4
Voyager 2023-5
GITM 2023-5
ToT 2023-5
RecAgent 2023-6
MIND2WEB 2023-6
ChatDev 2023-7
CO-LLM 2023-7
AgentSims 2023-8
Tool LLaMA
ToolBench 2023-7

1. Agent 关键组成

关键组成

- 规划 Planning + 记忆 Memory + 工具 Tools



1. 规划 Planning

- **规划**：一项复杂任务通常包括多个子步骤，Agent 需要提前将一项任务分解为多个子任务。
  - **子目标与分解（Subgoal and decomposition）**：Agent 将复杂任务分解为更小、更易于处理的子目标，从而实现对复杂任务的高效处理。
  - **反思与完善（Reflection and refinement）**：Agent 可以对历史的动作进行自我批评和自我反思，从错误中吸取教训，并为未来的步骤进行改进，从而提高最终结果的质量。

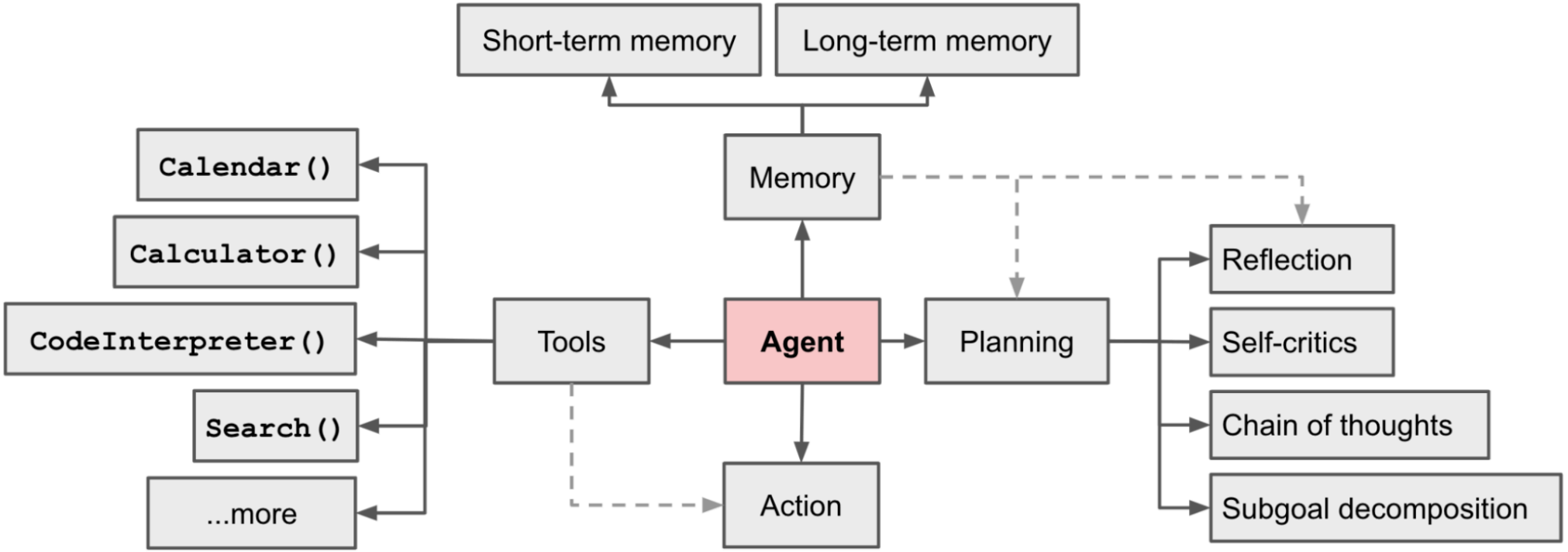
- **实现**：通过prompt engine来引导 LLM 实现规划（即步骤分解）。

2. 记忆 Memory

- **短期记忆（Short-term memory）**：所有上下文学习（In-context Learning），都是利用模型的短期记忆来学习。
  - 实现上主要利用 Prompt Engineering 。
- **长期记忆（Long-term memory）**：为 Agent 提供长时间保留和回忆信息的能力，这个时候需要借助外部向量存储和快速检索来实现
  - 实现上，主要利用向量数据库。

3. 工具 Tool

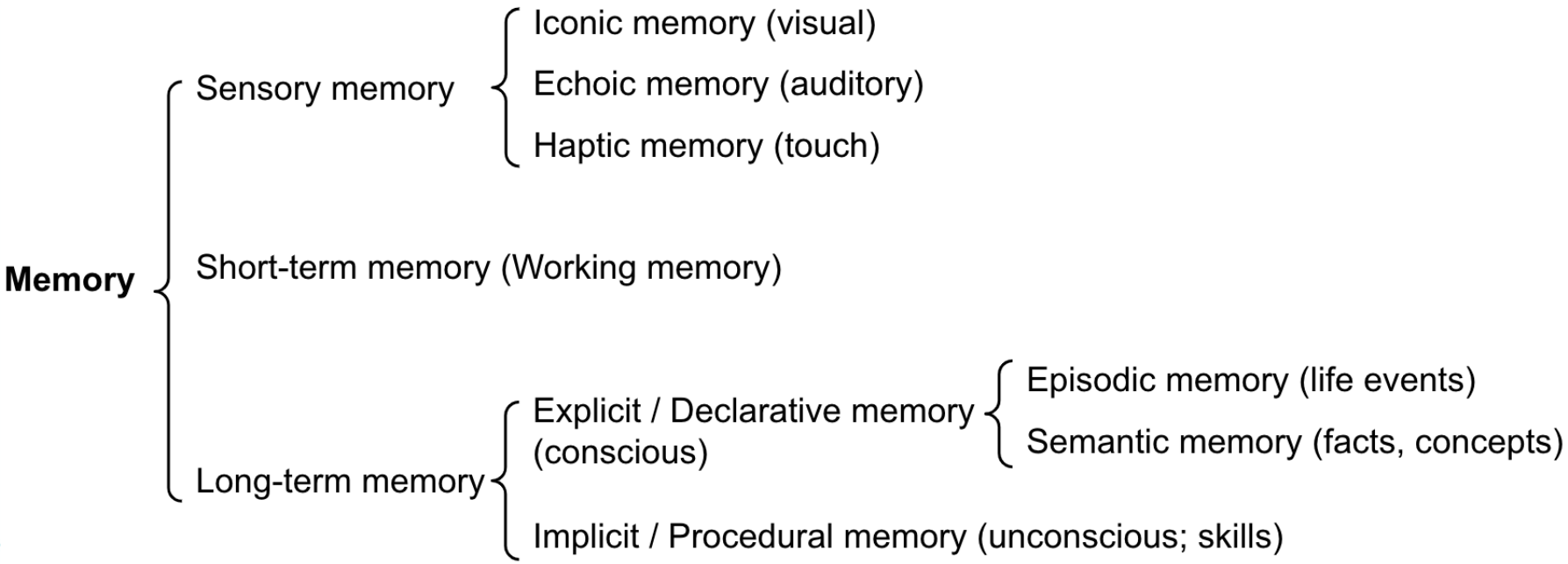
- Agent 会调用外部提供好的 API，补充 LLM 输出中缺失的额外信息，包括当前状态信息、具体的代码执行能力、访问专有信息源等，都需要借助外部的工具组件。



2. 记忆 Memory

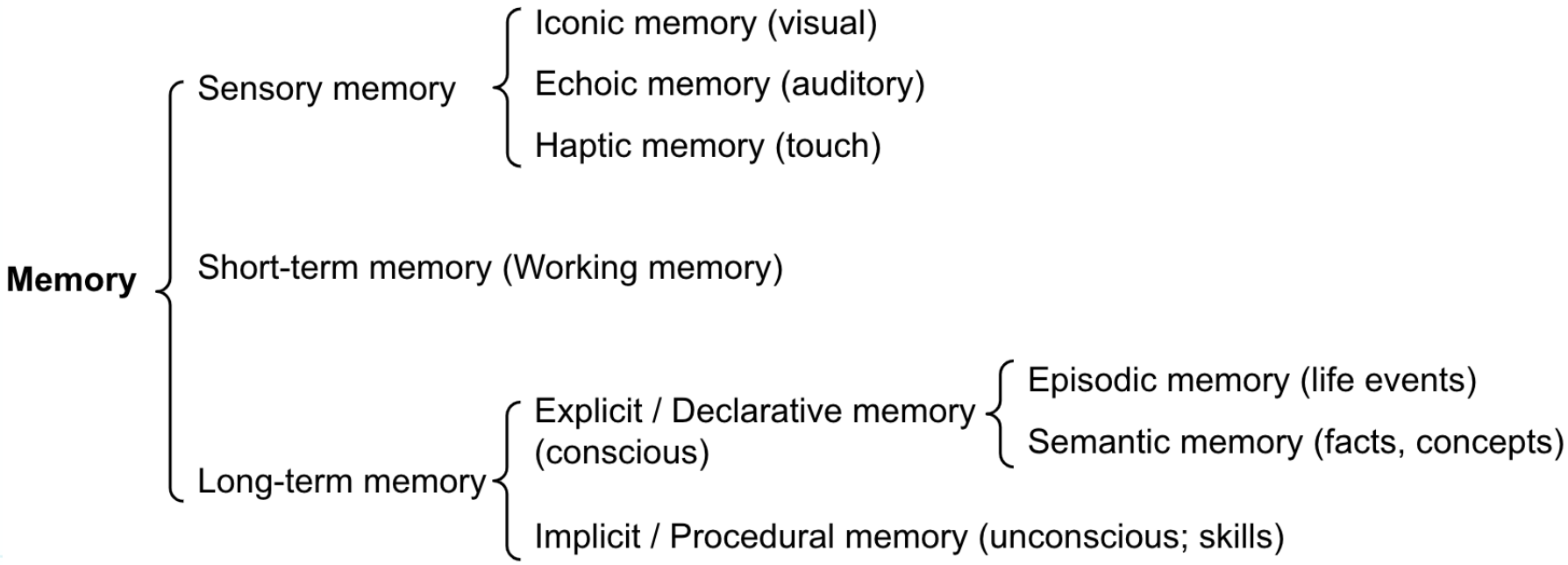
Human 记忆

- **感知记忆（Sensory Memory）**：记忆早期阶段，在原始刺激结束后保持对感官信息（视觉、听觉）的印象。
  - **特点**：持续时间短，子类包括图像记忆（视觉）、回声记忆（听觉）和触摸记忆（触感）。



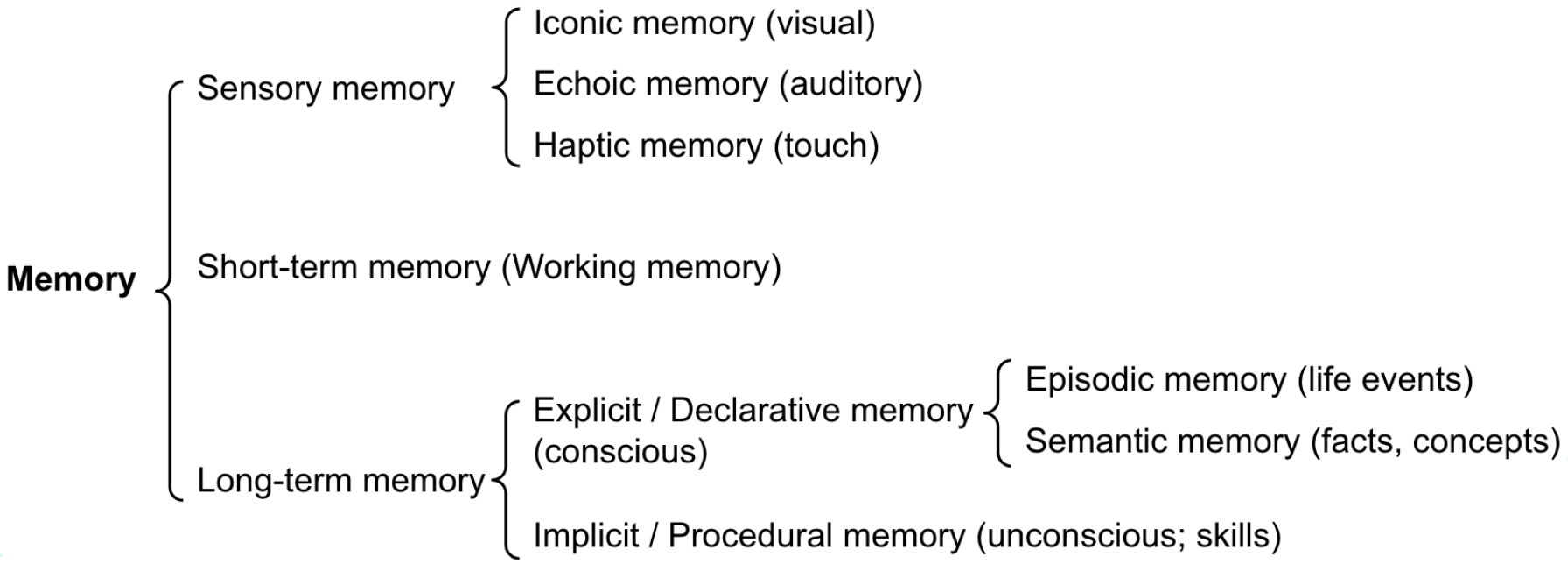
Human 记忆

- **短期记忆（STM，Sort-term Memory）**：短期记忆存储目前所知道的信息，以及执行复杂认知任务（如学习和推理）所需要的信息。
  - **特点：**短期记忆持续时间较短（1~2 周）



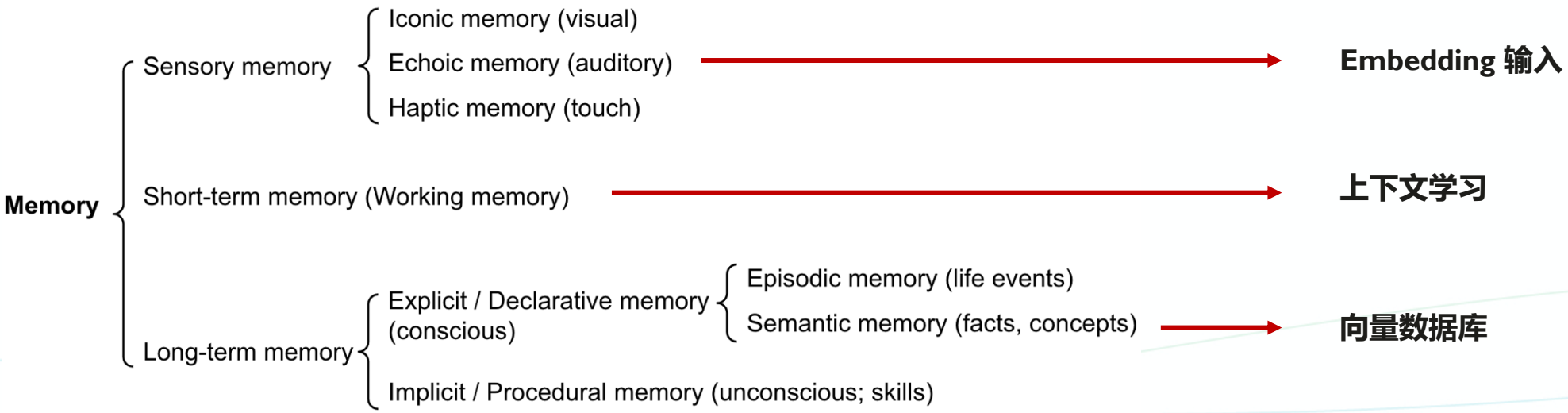
Human 记忆

- **长期记忆（Long-term Memory，LTM）**：将信息长时间存储，从几天到几十年不等。
  - 显式记忆（Explicit）：对事件的记忆，可以有意识地回忆起来的记忆，如记得小时候我喜欢吃奶瓶；
  - 隐式记忆（Implicit）：无意识的记忆，涉及自主执行的技能和惯例，如黄老师傅学会开车；



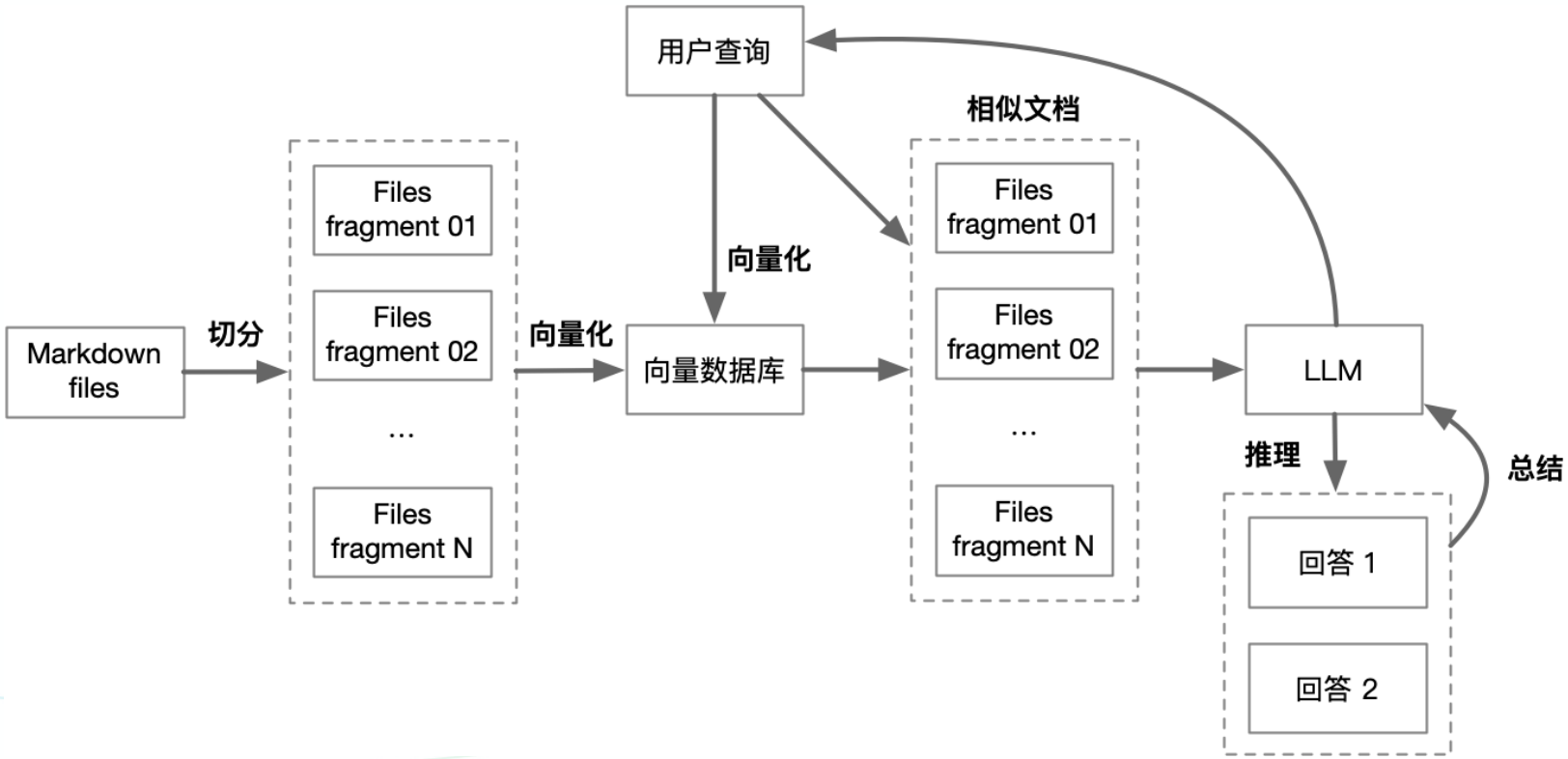
Human 记忆跟 LLM 结合

- First of all，人类擅长拥有不同类型记忆能力，而 LLM 很难理解新概念或者少量新数据。
  1. 感知记忆可以作为 LLM 或者多模态的 Embedding 输入表示（包括文本、图像等）。
  2. 短期记忆 STM 使用 In-context Learning ，受到 Transformer 有限 Seq Len 长度限制（2K to 8K）。
  3. 长期记忆 LTM 借助外部向量存储，Agent 可以快速查询、快速检索，从而进行访问。



LLM 长期记忆：语义搜索 dataset

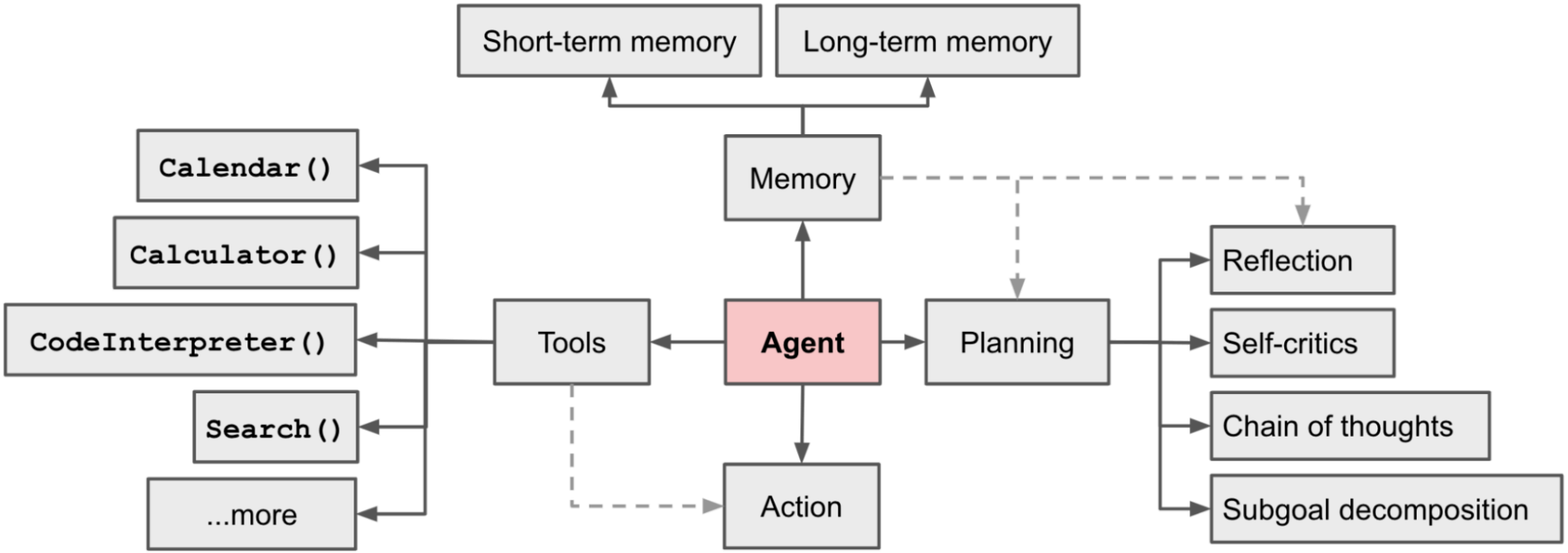
- 1）使用 Embedding 将记忆文本转化为向量 Vector；2）跟模型的交互信息同样通过 Embedding 转化为Vector，通过计算相似度来找到相似的记忆文本；3）将记忆文本拼接到 Prompt 作为模型输入。



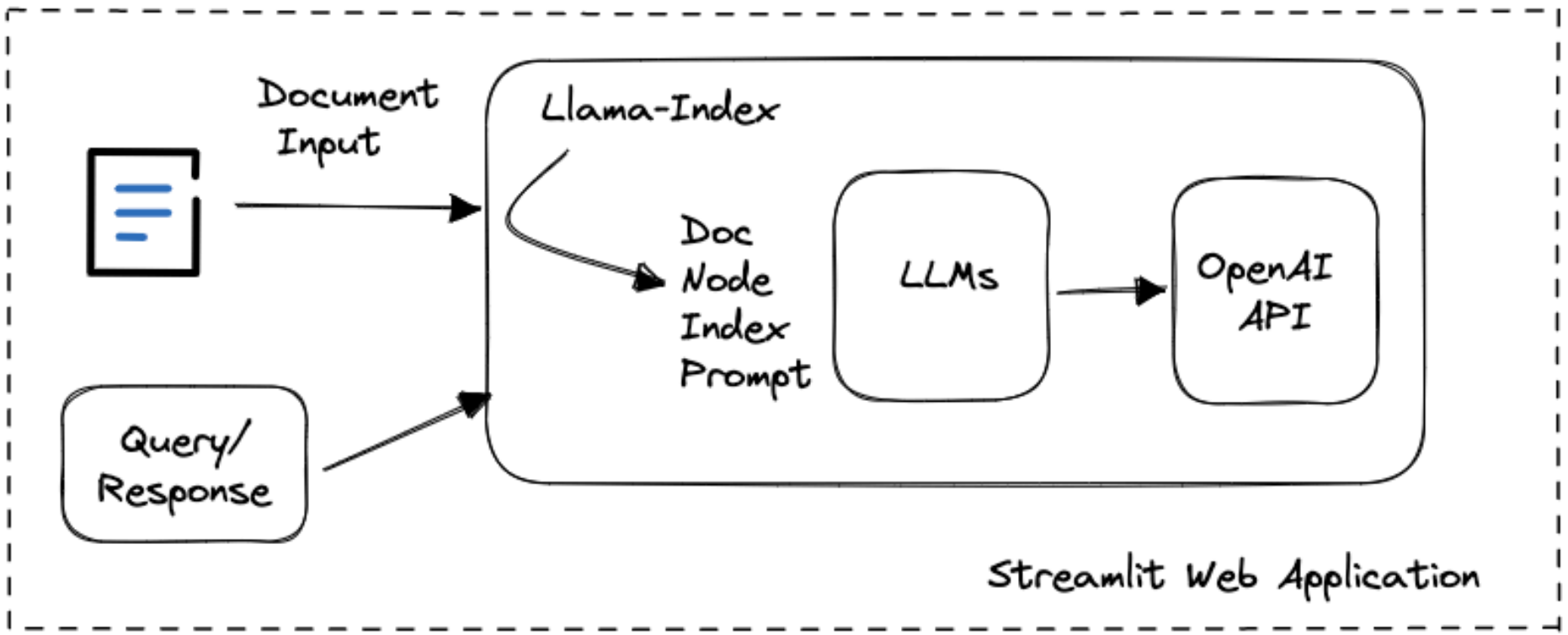
3. Tool Use

工具使用

Tools and Action

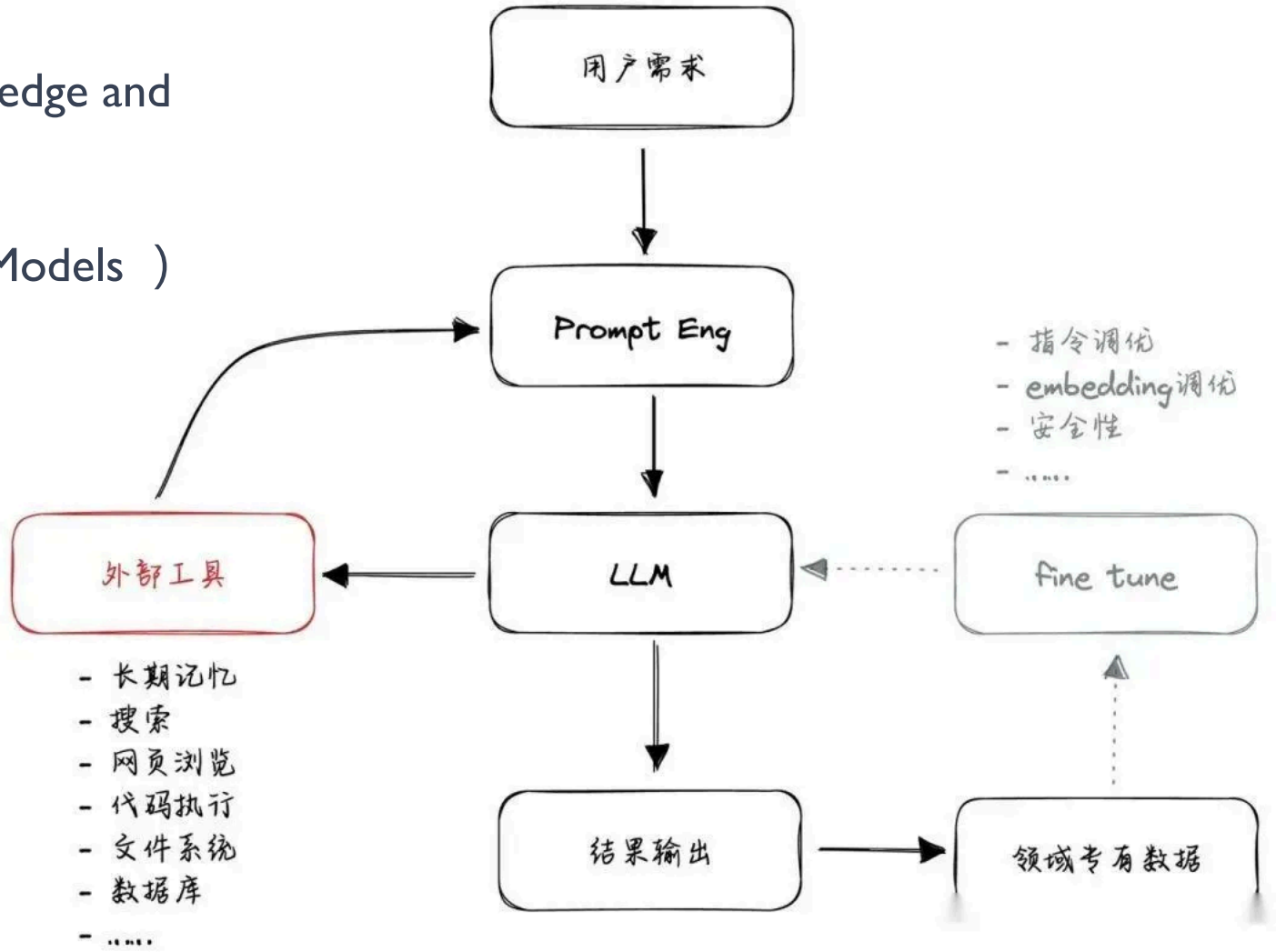


工具的作用


Document
Input
Llama-Index
Doc
Node
Index
Prompt
LLMs
OpenAI
API
Query/
Response
Streamlit Web Application

工具的作用

- MRKL（Modular Reasoning，Knowledge and Language）
- TALM（Tool Augmented Language Models ）



1. MRKL

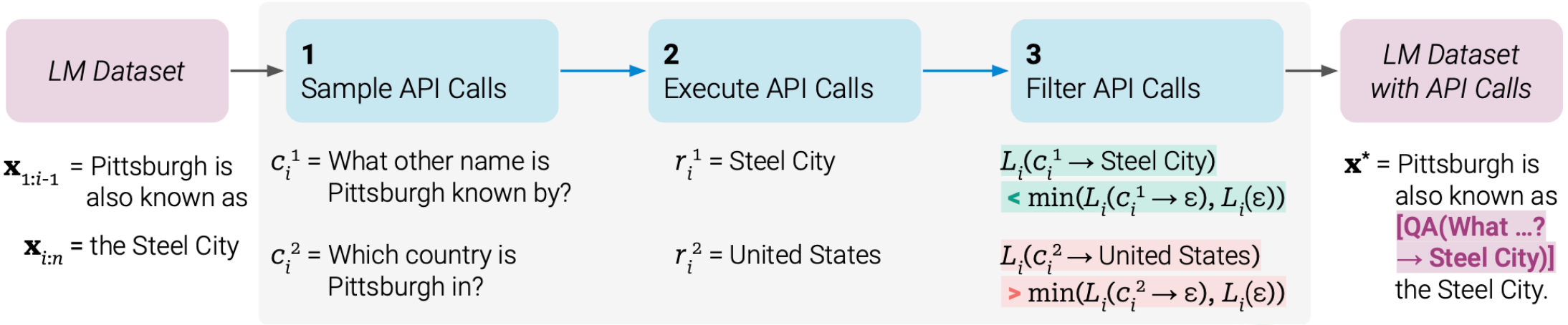
- **MRKL**（Karpas et al. 2022， Modular Reasoning, Knowledge and Language ），用于自主代理的神经符号结构。
  - 包含一组「专家」模块和一个用作「路由 Router 」的 LLM，通过路由查询到最合适的专家模块（Like MOE？）
  - 每个「专家」模块可以用神经网络表示，也可以使用符号模型表示，例如数学计算器 API、天气 API、Python 执行器，而我最常用的仅仅是闹钟。。。

1. MRKL 例子

- 小艺小艺，今天天气好伐？
- 小艺小艺，帮侬调个闹钟

2. TALM

- **TALM**（Tool Augmented Language Models，Parisi et al. 2022）通过微调 LLM 来学习使用外部工具 API。
  - 数据集根据新增的API调用注释，来提高 LLM 输出的质量来进行扩展。



2. TALM 例子

- HuggingGPT（Shen et al. 2023），利用 ChatGPT 作为任务规划器，根据模型描述选择 HuggingFace 平台上可用的模型，并根据执行结果总结 Feedback。
- ChatGPT 插件和 OpenAI API 函数调用，其中工具 API 集合可以由其他开发人员提供（如插件）或自定义（如函数调用）。

4. 规划 Planning

规划 Planning

- **规划：**一项复杂任务通常包括多个子步骤，Agent 需要提前将一项任务分解为多个子任务。
  - **子目标与分解（Subgoal and decomposition）**：Agent 将复杂任务分解为更小、更易于处理的子目标，从而实现对复杂任务的高效处理。
  - **反思与完善（Reflection and refinement）**：Agent 可以对历史的动作进行自我批评和自我反思，从错误中吸取教训，并为未来的步骤进行改进，从而提高最终结果的质量。

- **实现：**通过prompt engine来引导 LLM 实现规划（即步骤分解）。

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem