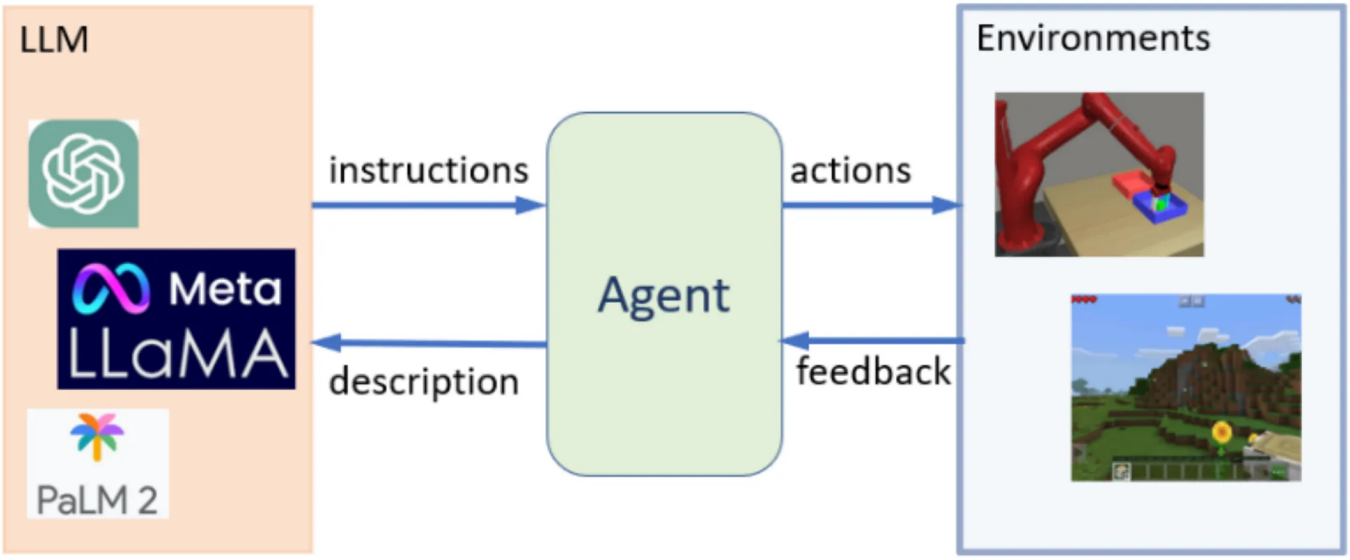
LLM Agent（智能体）

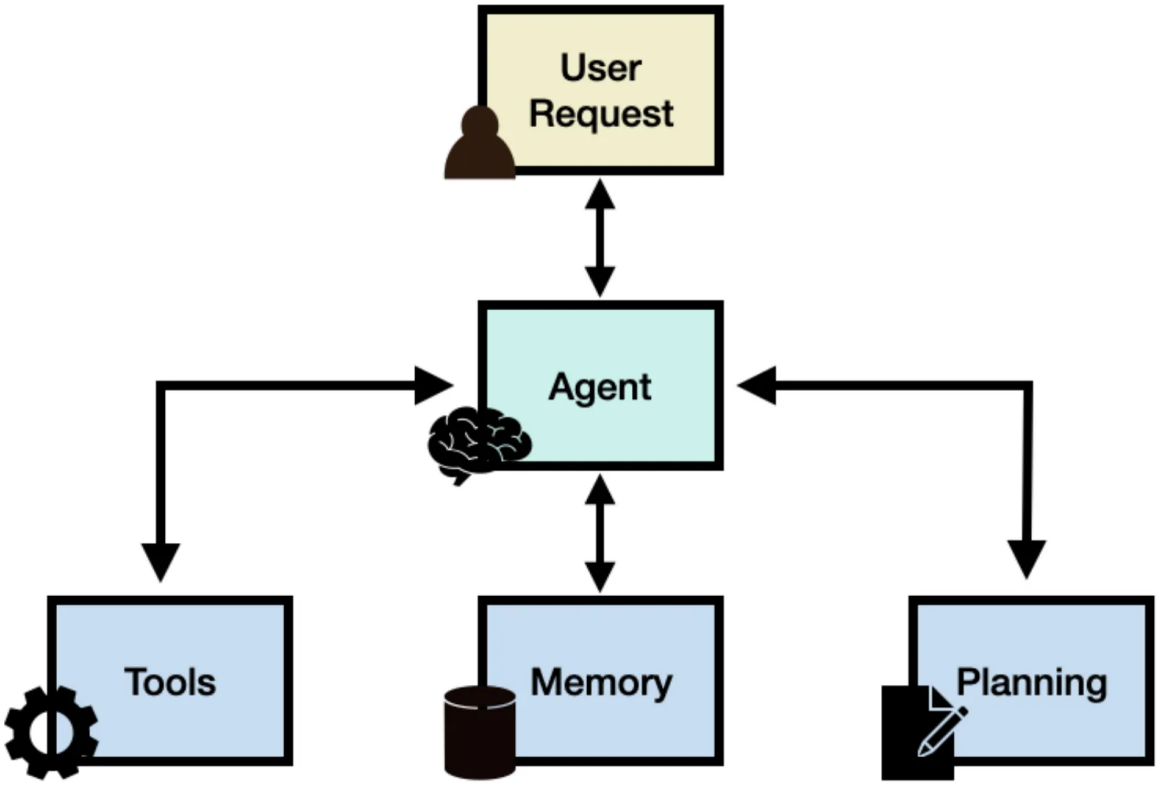
一、LLM Agent

大模型Agent是一种构建于大型语言模型（LLM） 之上的智能体，它具备环境感知能力、自主理解、 决策制定及执行行动的能力。

Agent是能够模拟独立思考过程，灵活调用各类工具，逐步达成预设目标。在技术架构上，Agent从 面向过程的架构转变为面向目标的架构，旨在通过感知、思考与行动的紧密结合，完成复杂任务。



大模型Agent由规划、记忆、工具与行动四大关键部分组成，分别负责任务拆解与策略评估、信息存储与回忆、环境感知与决策辅助、以及将思维转化为实际行动。



1. **规划**（Planning）：

定义：规划是Agent的思维模型，负责拆解复杂任务为可执行的子任务，并评估执行策略。

实现方式：通过大模型提示工程（如ReAct 、 CoT推理模式）实现，使Agent能够精准拆解任务，分步解决。

2. **记忆**（Memory）：

定义：记忆即信息存储与回忆，包括短期记忆和长期记忆。

实现方式：短期记忆用于存储会话上下文，支持多轮对话；长期记忆则存储用户特征、业务数据等，通常通过向量数据库等技术实现快速存取。

3. **工具**（Tools）：

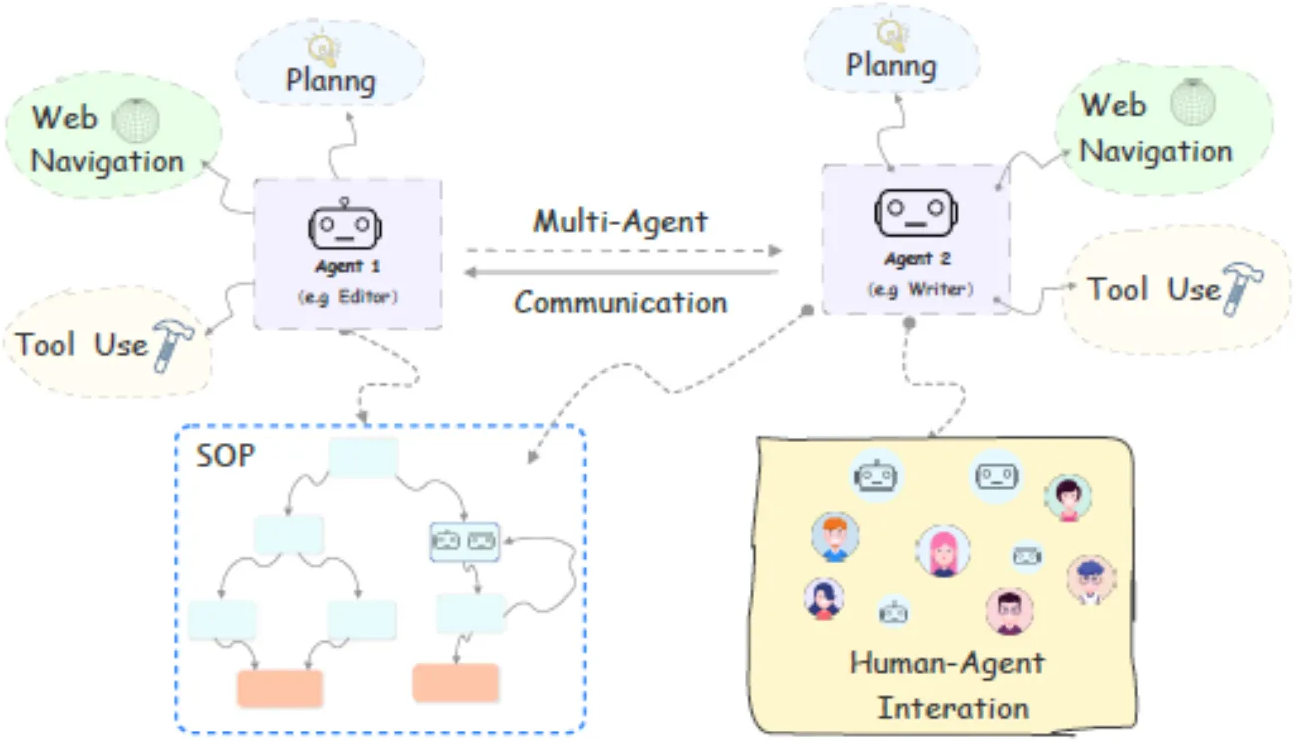
定义：工具是Agent感知环境、执行决策的辅助手段，如API调用、插件扩展等。

实现方式：通过接入外部工具（如API 、插 件 ）扩展 Agent的能力，如 ChatPDF解析文档 、Midjourney文生图等。

4. **行动**（Action）：

定义：行动是Agent将规划与记忆转化为具体输出的过程，包括与外部环境的互动或工具调用。

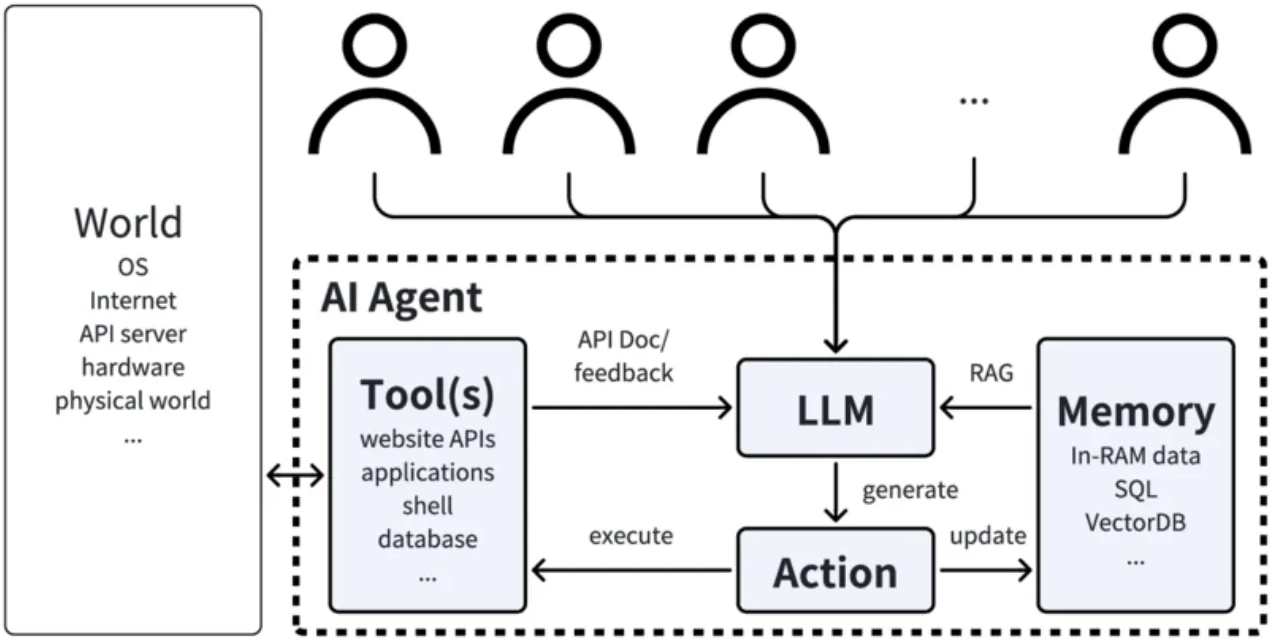
实现方式： Agent根据规划与记忆执行具体行动，如智能客服回复、 查询天气预报 、AI机器人抓起物体等。



二、LLM Agent + RAG

RAG技术为LLM Agent提供了额外的知识来源 。传统的LLM虽然能够从大规模文本数据中学习到丰富的语言知识和模式， 但它们在处理特定领域或需要专业知识的问题时可能表现不足。

通过引入RAG ， LLM Agent能够在需要时查询外部知识库，如专业数据库、 学术论文、行业报告等，从而增强其知识广度和深度。


World
OS
Internet
API server
hardware
physical world
...
AI Agent
Tool(s)
website APIs
applications
shell
database
...
API Doc/
feedback
LLM
RAG
Memory
In-RAM data
SQL
VectorDB
...
generate
execute
Action
update