

# AI 应用开发

adamanteye

## 1. AI Agent

### 1.1. 缓存

AI Agent 的输入 Token 与输出 Token 常常是不成比例的,在 Manus 中,平均输入对输出的比率约为 100:1. 因此,每次输入应当为仅追加,并且保持提示词稳定.

### 1.2. MCP

MCP (Model Context Protocol)赋予了大模型调用系统工具,外部 API,本地数据的能力.

传统的 Function Calling 将大模型可调用的工具放在 system prompt 里面,随 user prompt 发给大模型,大模型的输出被解析为具体的调用.

而在 MCP 的框架里,工具和资源由 MCP Server 管理,MCP Host (AI Agent)与大模型以及 MCP Server 交互.因此,MCP 赋予了大模型协调工具,实现流程自动化的能力.

### 1.3. 概念

**Host** LLM applications that initiate connections

**Client** Connectors within the host application

**Server** Services that provide context and capabilities

## 2. RAG

检索增强生成,即 RAG (Retrieval-Augmented Generation),是一种使用来自私有或专有数据源的信息来补充文本生成的技术.

Embedding,即将限制长度的输入转为高维向量的技术,以及向量数据库,是实现文档入库以及查询的核心.

## 3. 模型

### 3.1. API 调用

- [Portkey-AI/gateway: A blazing fast AI Gateway with integrated guardrails](#)

### 3.2. Embedding

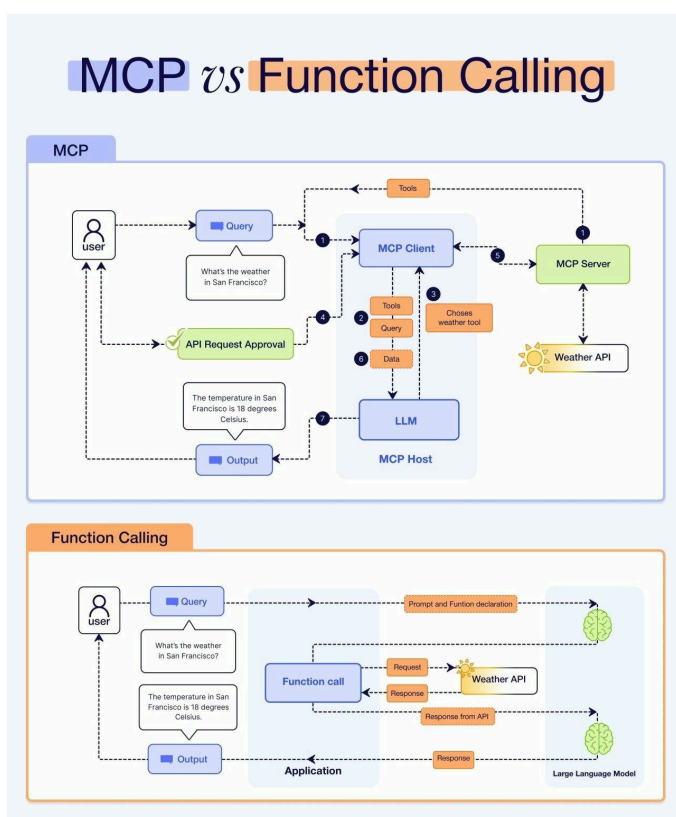


图 1 MCP 流程

在基于 MCP 构建应用的过程中,有以下难题需要解决:

- 网页浏览
- 沙箱
- 用户记忆
- 数据库