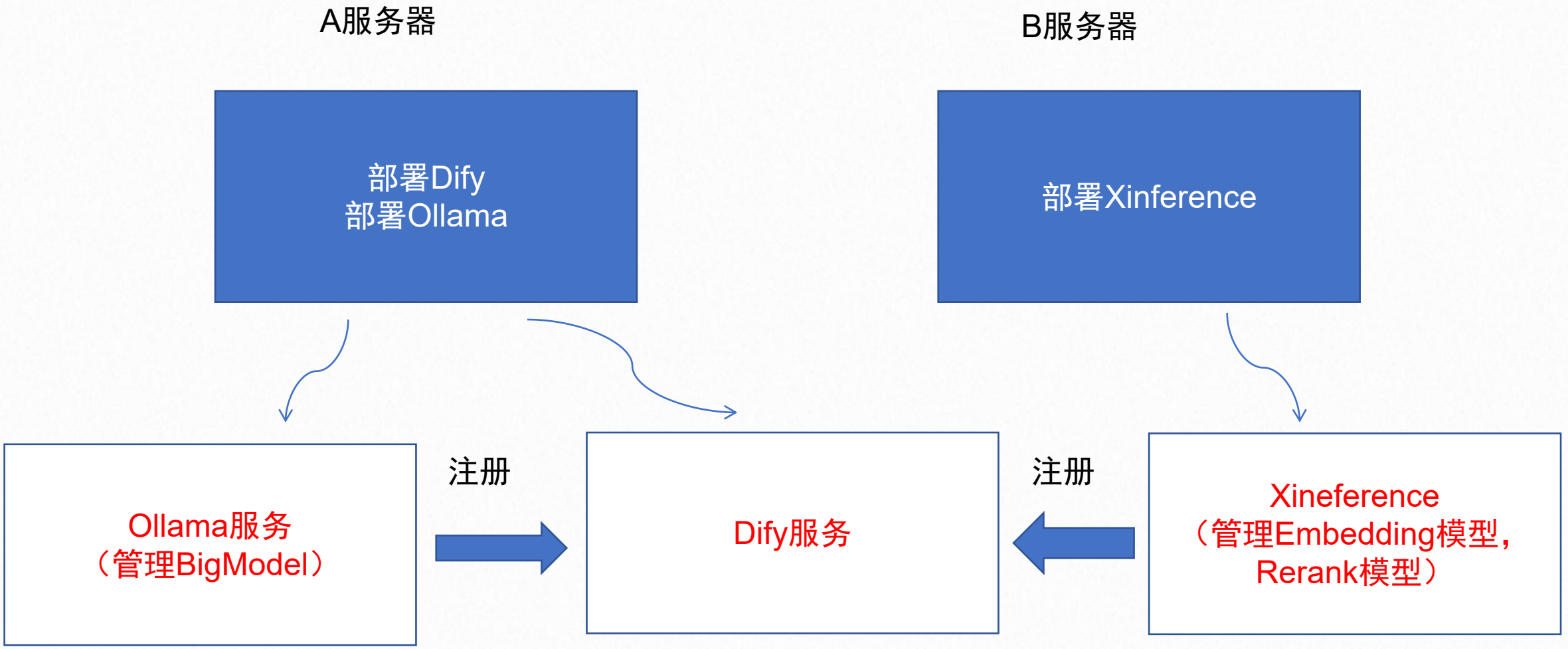
Dify智能应用开发实战

目录

1 Dify平台私有化部署

01、Dify平台私有化部署



02、大模型高性能管理平台

01

Ollama

Ollama 是用于私有化部署和运行大型语言模型的工具，支持 通义千问、小羊驼 等多种模型，私有化部署使用户没有网络连接的情况下也能使用这些先进的人工智能模型。

02

Xinference

xinference 是一个强大且通用的分布式推理框架，也可以用于私有化部署和运行大语言模型，通过 xinference 可以简化各种AI模型的运行和集成。

03

LocalAI

LocalAI是一个与OpenAIAPI兼容的本地RESTAPI，它使用C++绑定优化速度，支持在消费级硬件上运行多种大模型。

04

FastChat

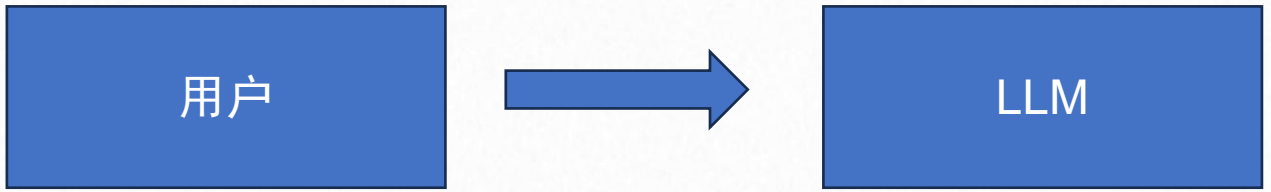
FastChat 是一个用于训练、部署和评估大模型的开源库。

2 Dify案例实践之对话应用构建

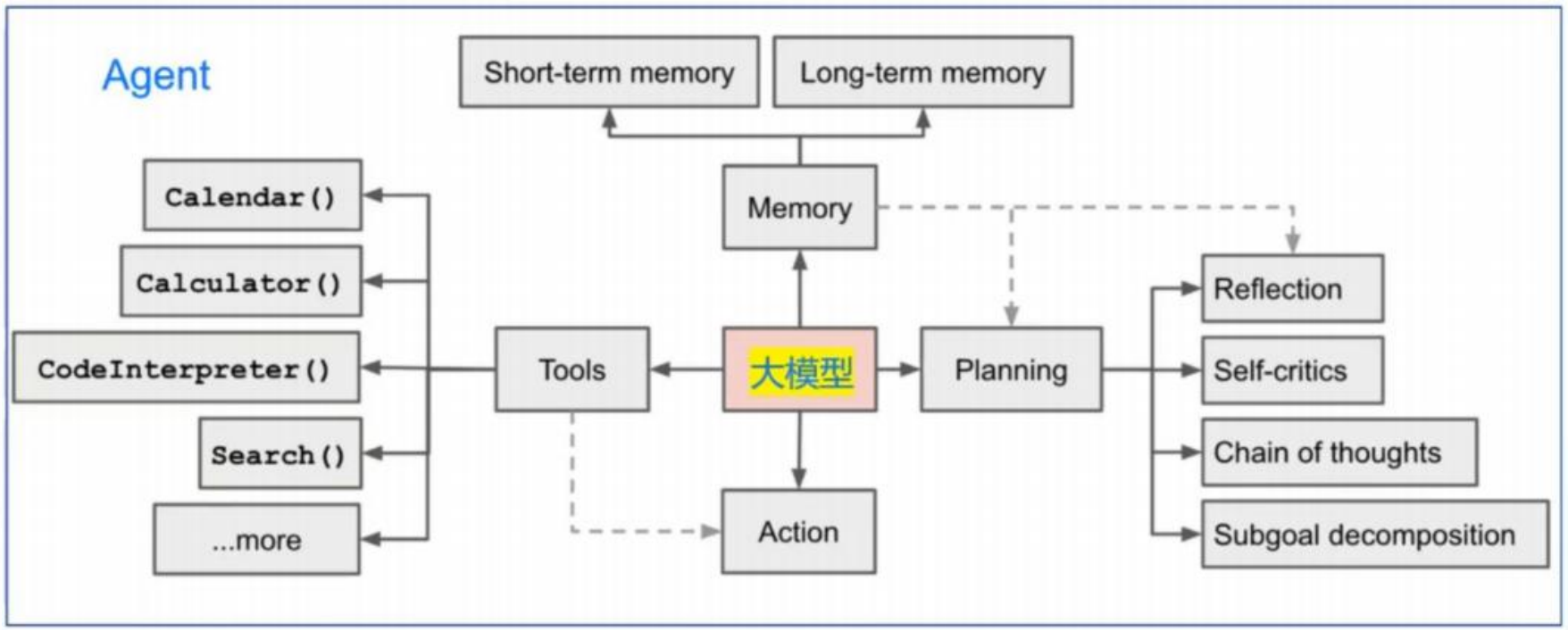
3 Dify案例实践之智能体应用构建

4 Dify案例实践之工作流应用构建

什么是Agent(智能体)?



问题过于复杂
无法解决用户问题



面向Agent编程

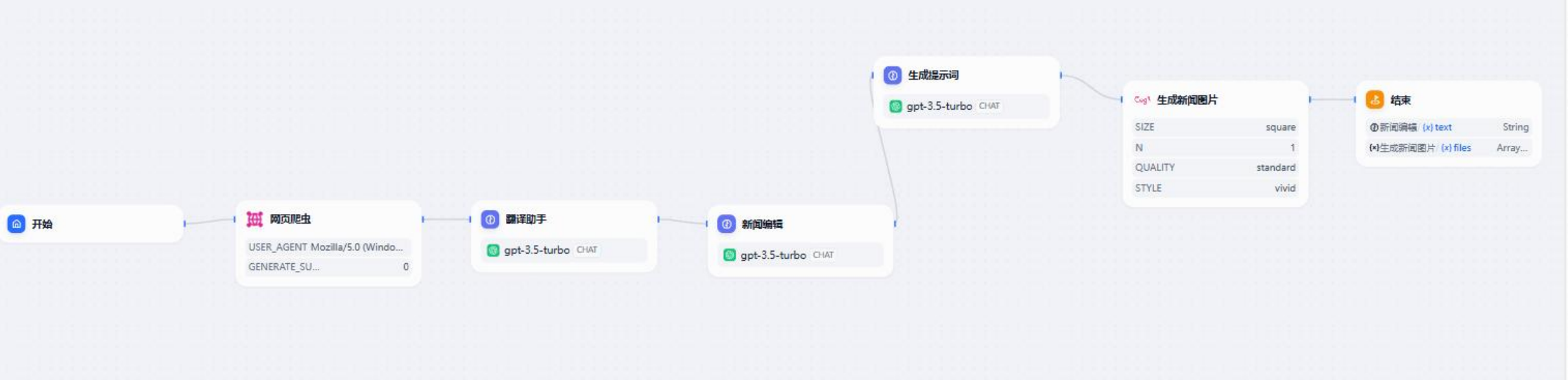
比较灵活

解决复杂问题

什么是WorkFlow(工作流)

用户 → LLM

问题过于复杂
无法解决用户问题

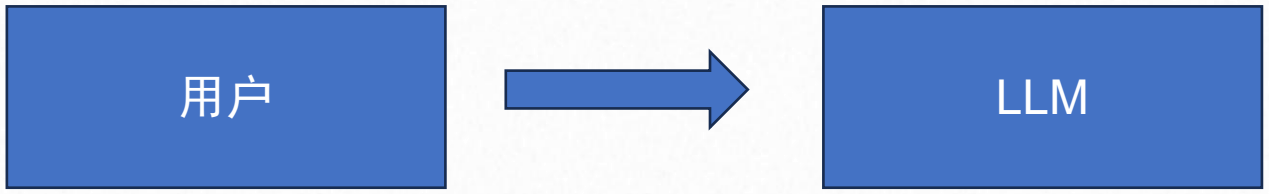


面向过程编程

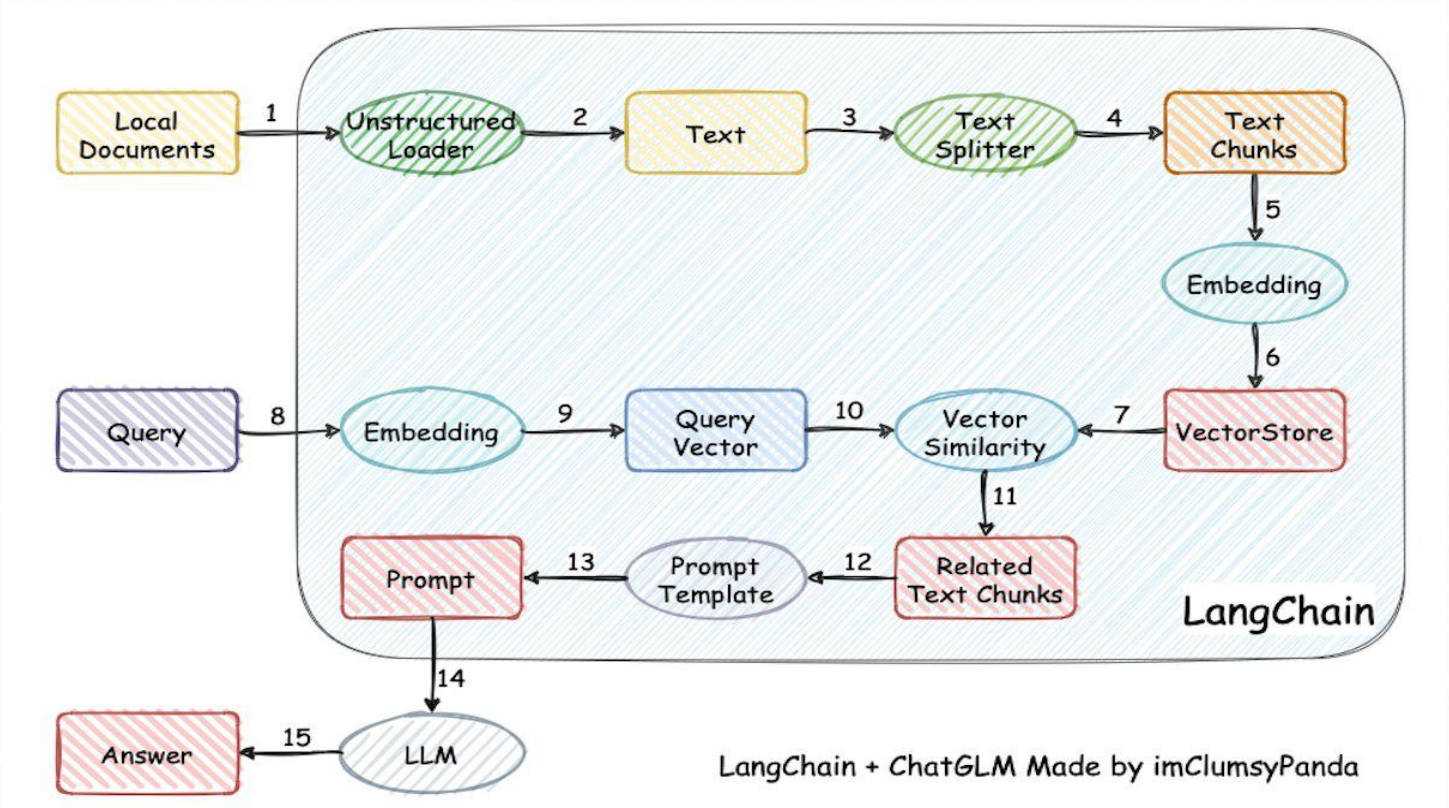
效果稳定

解决复杂问题

什么是RAG(检索增强生成)?



知识不足
无法回答用户问题



面向过程编程

效果稳定

解决大模型知识不足问题

大模型架构新范式的优缺点

	优点	缺点
Agent	动态规划 灵活	缺乏稳定性
Workflow	静态规划 稳定性高	缺乏灵活性
RAG	静态规划 效果稳定	缺乏灵活性

5 Dify案例实践之私有知识库构建