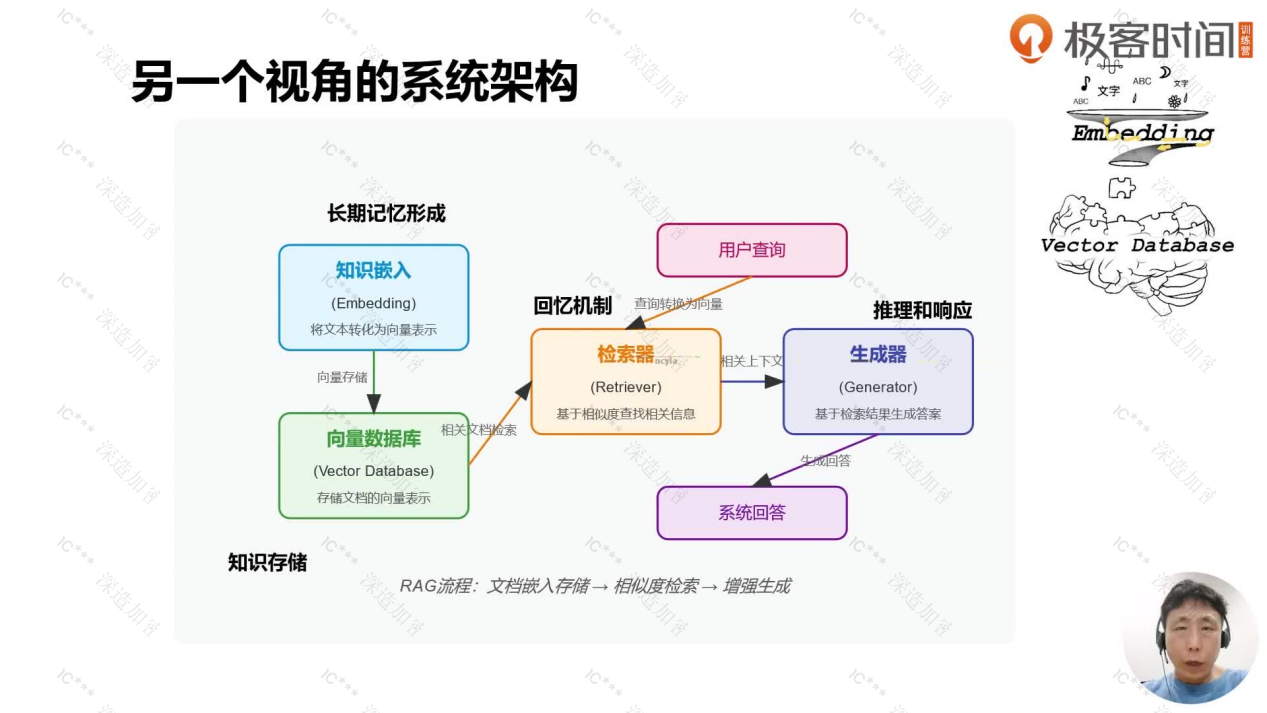
# RAG基础和项目落地

1. RAG三问
2. RAG整体框架和技术组件

1.1 AI时代的应用开发范式发展



1.2 RAG的基本框架

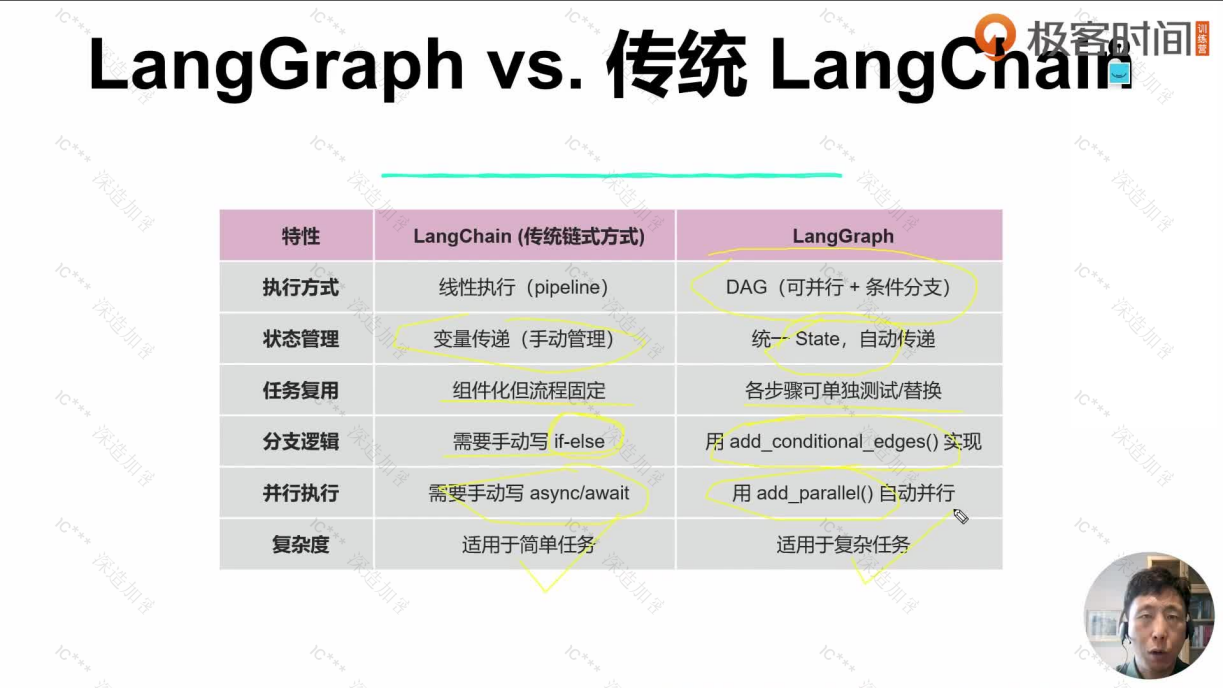


1.3 四大核心组件：知识嵌入、向量数据库、检索器、生成器(LLM为核心)，最重要是知识嵌入和检索器这两部分

2. 快速搭建RAG

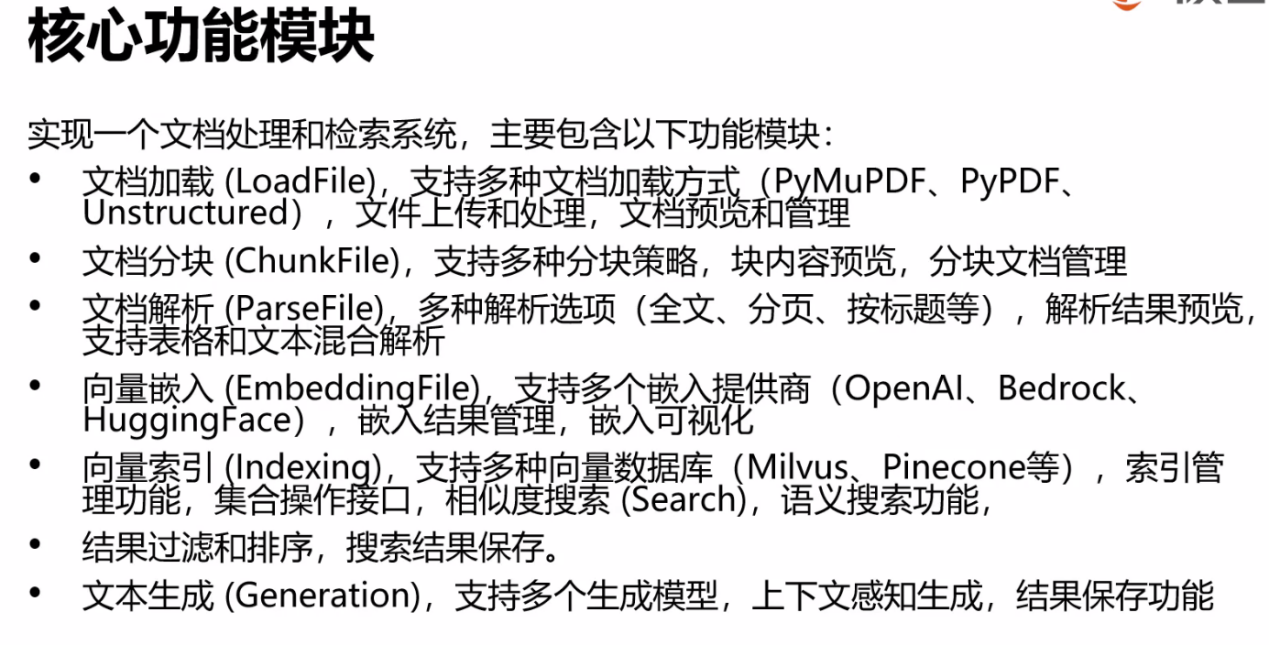
方法：Llmaindex、langchain、LCEL、LangGraph、手动搭建

对比：



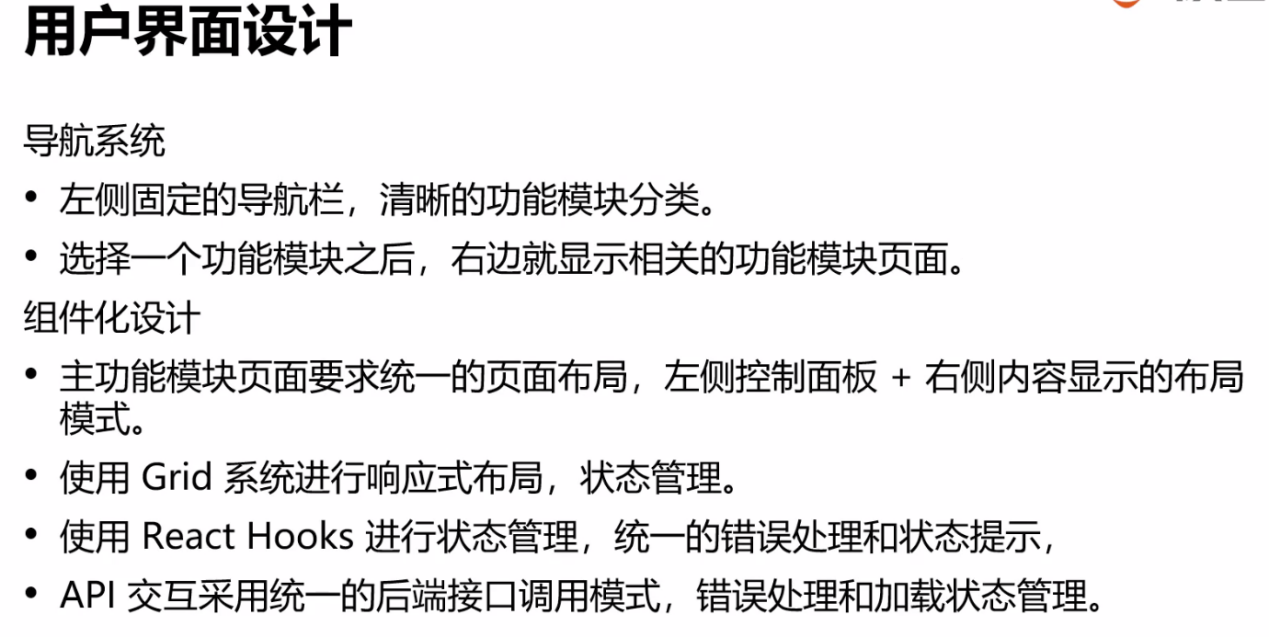
二. 构建RAG系统

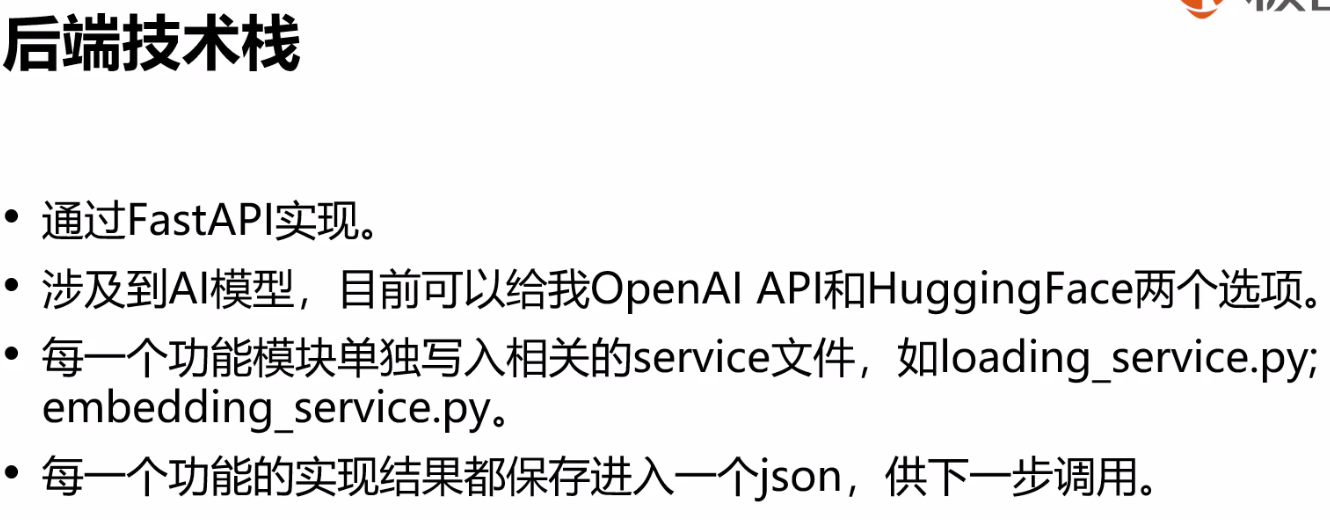
1. 核心功能



2. 前端架构



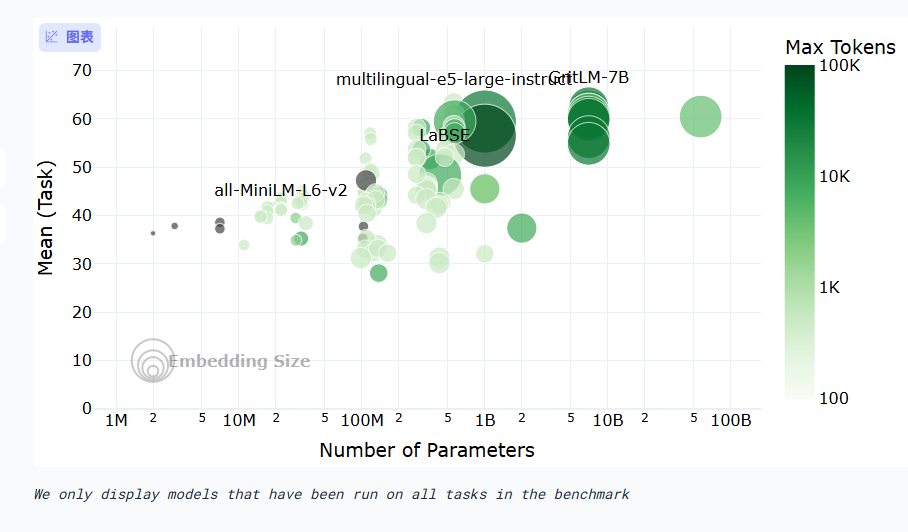


1. 后端架构

三. 项目---医疗系统

1. embedding模型选型

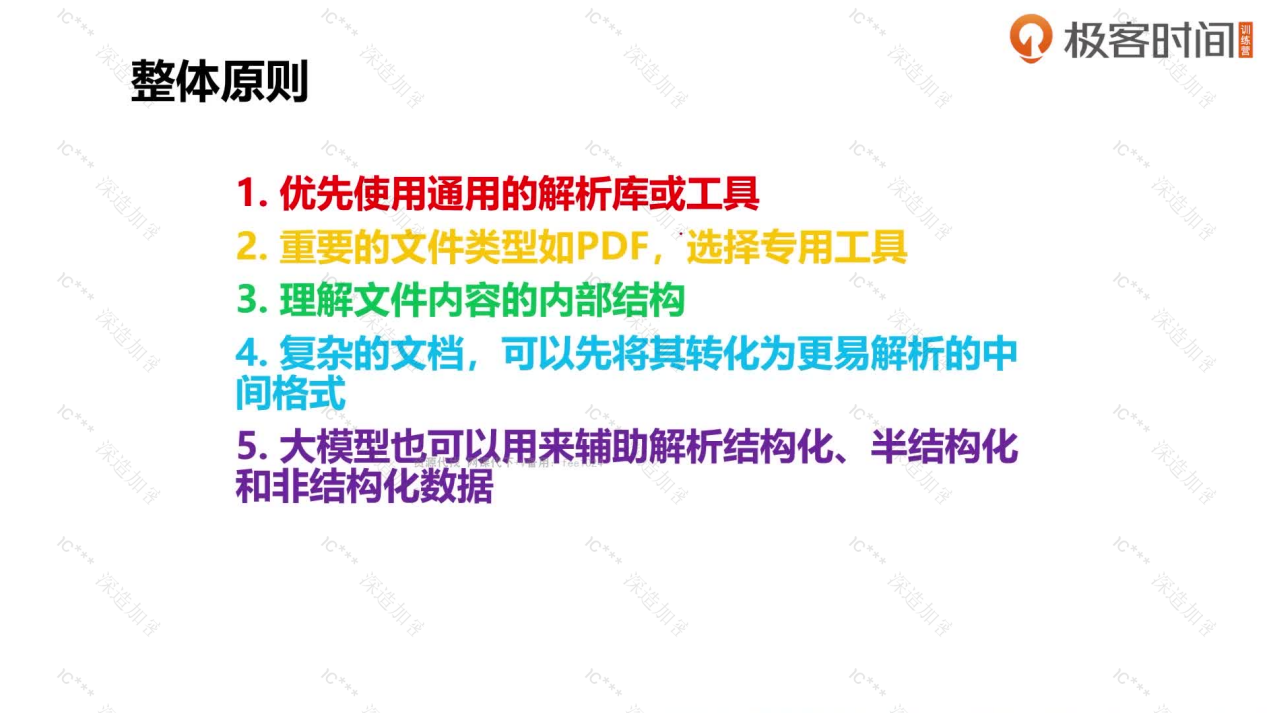
可到mteb查看效果较好的模型, 通常7B属于很大

https://huggingface.co/spaces/mteb/leaderboard 

越靠右越大, 越考上分数越高, 但目前效果最好的是openai的text-embeding-3-large收费模型

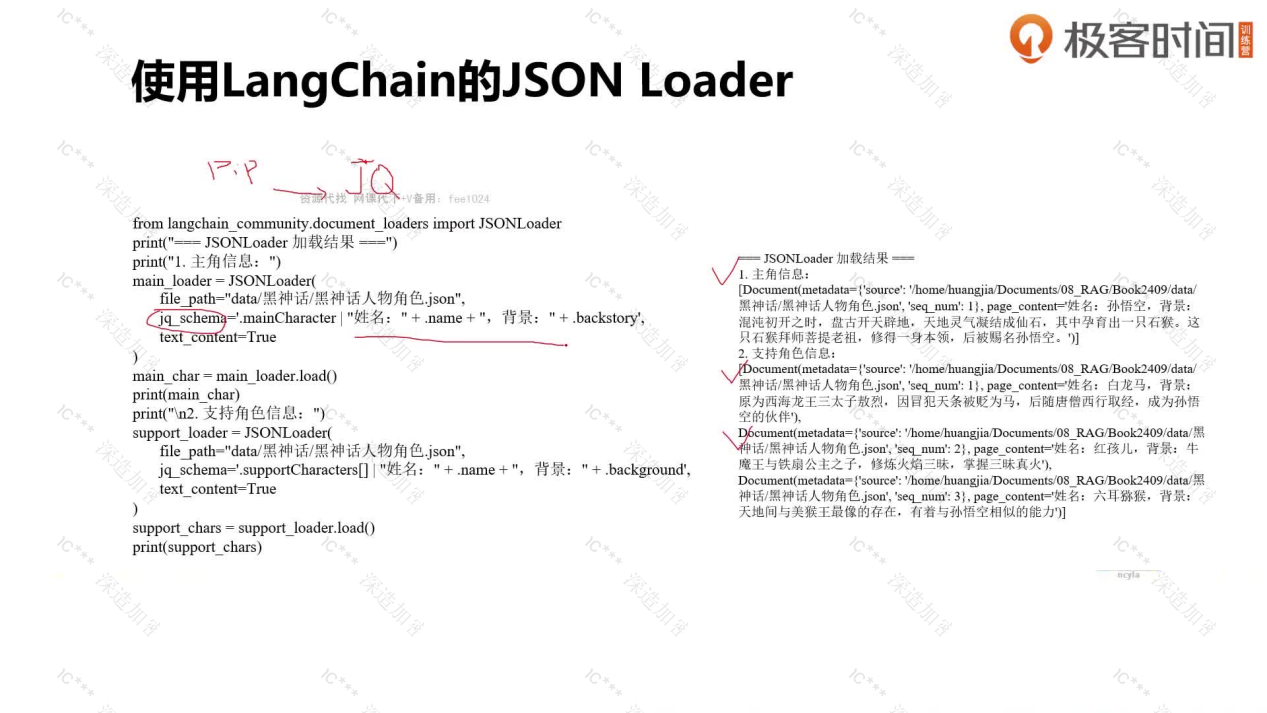
注: 获取不同嵌入模型的维度不需要手动制定, 可以直接嵌入一个字符串, 从结果获取

1. 组件——数据导入
2. 整体原则



1. 结构化提取

2.1 Json可设定jq\_shema进行一定格式的提取



2.2 UnstructuredLoader可对url或html等文件进行提取，能获取元素间的关系

UnstructuredMarkdownLoader可对markdown进行提取，能获取元素间的关系

上述注意处理父子元素等，放在一起保存

1. 图文数据的提取
   1. UnstructuredImageLoader可直接OCR识别图片中的文字，但效果一般
   2. unstructured.partition.ppt可读取PPT并识别文字
   3. from pdf2image import convert\_from\_path将PDF转成图片后，再调用多模态大模型进行图片含义的理解
2. PDF解析(重点)

主要解析方法：

基于规则(pyPDF, pyMUPDF)

基于深度学习(pytesseract+pdf2image属于OCR)

基于大模型的理解

unstructured属于规则和深度学习的结合，更实用

4.1 pdf转markdown: Marker，MinerU和LlamaParse均是通过深度学习OCR等方法