腾讯云开发者社区 | 腾讯云向量数据库 Tencent Cloud VectorDB | Andon

腾讯云向量数据库

RAG七天入门课 第一节

“消灭”LLM幻觉的利器
— RAG介绍

腾讯云向量数据库产品经理

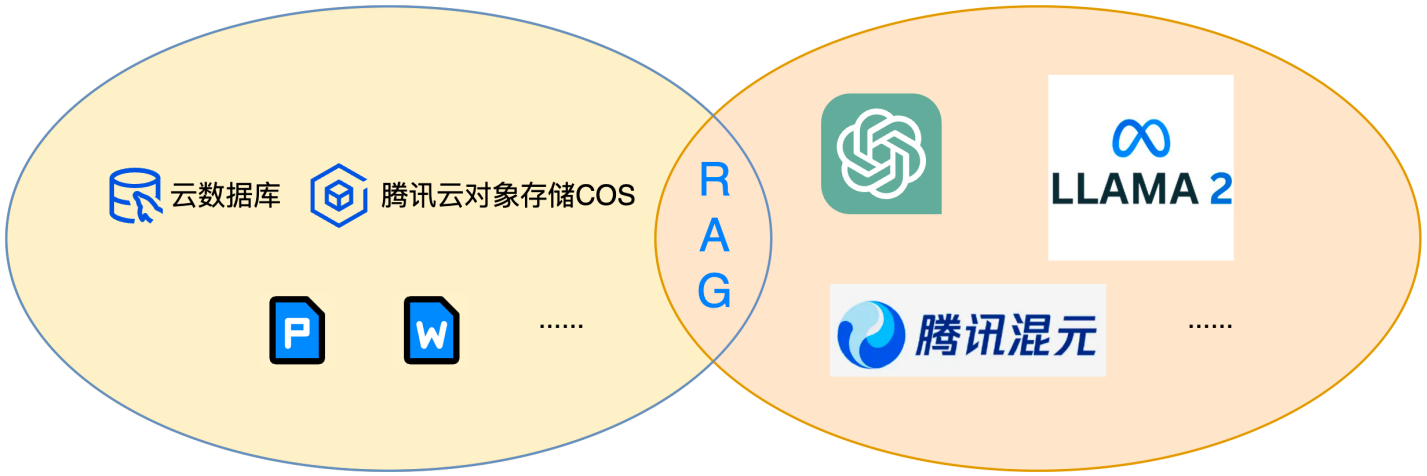
熊 鑫

BBTalk

第一节

“消灭”LLM幻觉的利器 - RAG介绍

Why RAG ?



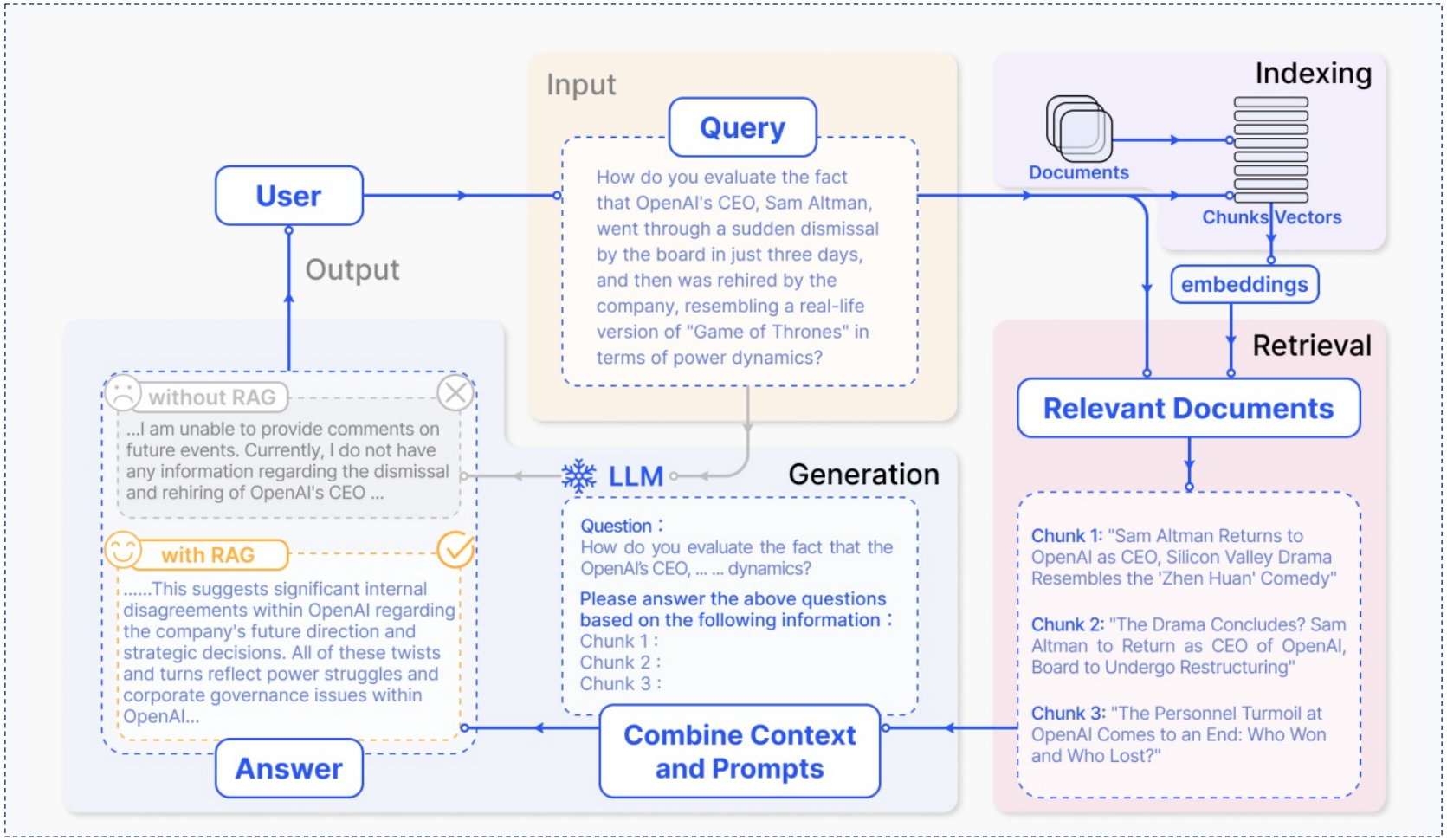
企业私域数据&实时更新的公域数据

预训练大语言模型

通用大语言模型LLM已火遍全球，RAG结合LLM帮助您构建基于私有文档、专业领域知识、实时信息的Chatbot，为您的公司团队和客户提供更优质的服务。

- 在过去几乎无法完成
- 没有RAG很难做到

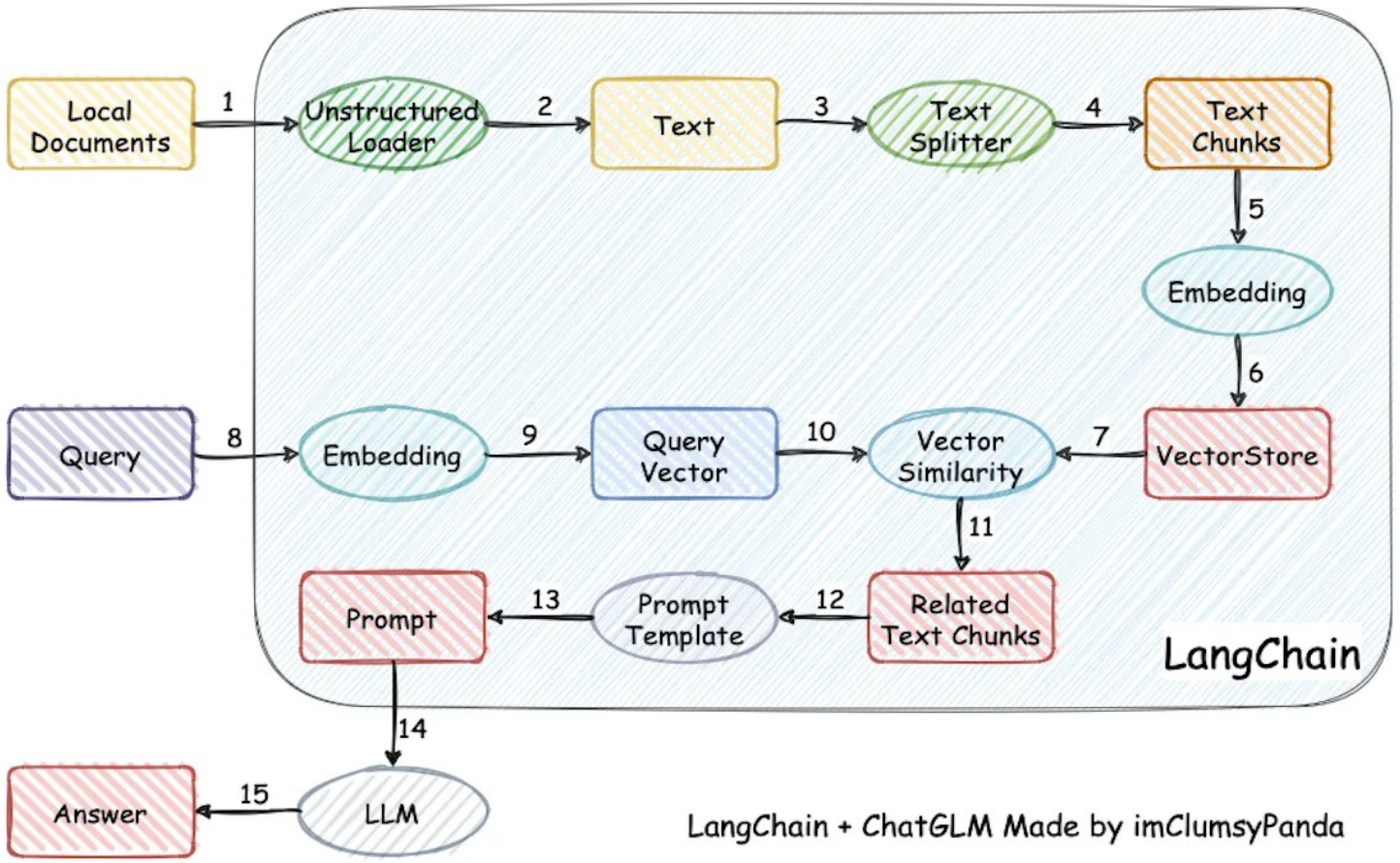
What is the RAG ? (Retrieval-Augmented Generation)



主要步骤

- 知识切片成Chunk
- 向量化Chunk入库
- Query检索知识Chunk
- 构建Prompts
- 调用LLM生成回答

基于开源构建RAG应用



优势

- 快速构建Demo
- 快速理解RAG
- 社区支持

痛点

- 投入大
- 效果差
- 调优难

Bad case 1

Eample：

相关知识：“...A产品的分析报告，会分析近30天的数据分析结果...”

用户问：“请问A产品分析报告多久分析一次？”

LLM回答：“XX产品每30天分析一次”

分析：

- 用户的问题在知识中未明确提及到
- 返回了有一定相似度的信息，但本身与用户问题不直接相关
- LLM未完全理解，开始“胡说八道”

Bad case 2

Eample：

- 期望知识，在【A课程文档】中：“...可能分级如下...其中XX等级适合XX-XX年龄的小孩...”
- 干扰知识，在其它文档中：“该课程适合3-7岁的小孩”、“可能适合6-8岁的女孩”
- 用户问：“请问A课程适合多大年龄的小孩？”
- LLM回答：“A课程适合3-7岁的小孩”

分析：

- 期望的A课程知识片段在内容中并未提及A课程，导致搜索分数不高
- 干扰知识中有明确的小孩年龄分布，导致其搜索排序效果更高
- LLM无法完全理解

Bad case 3

Eample：

- 用户问题1：“服务器连接不上了？应当如何解决？”
- 实际搜索结果：“连接服务器，步骤1：... 步骤2 ... 步骤3 ...”
- 用户问题2：“为什么内存变高了？”
- 实际搜索结果：“大Key处理方案...大Value处理方案...并发...”

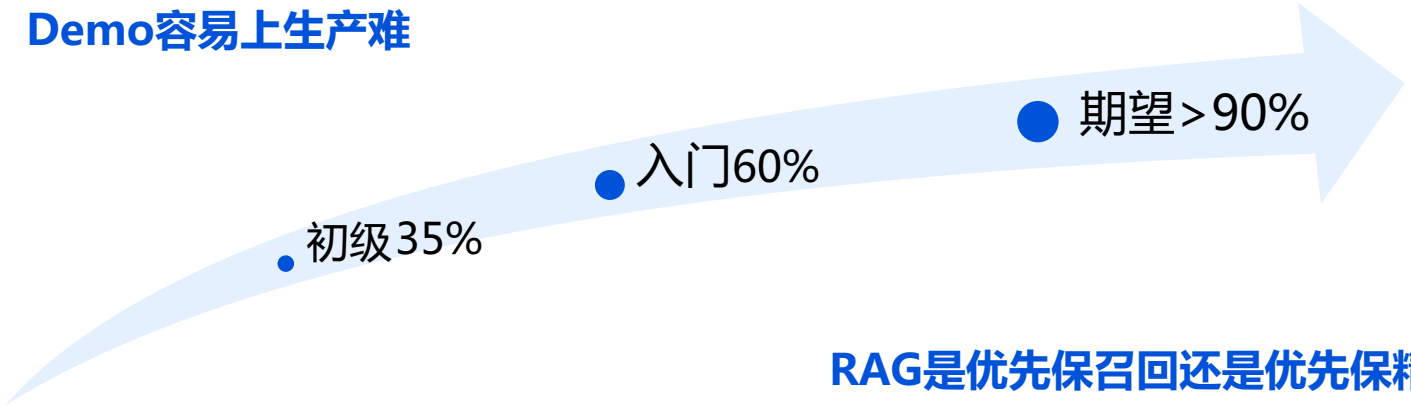
分析：

- 以上问题在知识中有“相似”但并非“直接切要害”的回答
- 第1个问题，应当有专门的QA文档引导文档，告知用户解决连接问题的排查
- 第2个问题，最佳的方式通过意图识别，路由到“诊断引擎”为用户排查问题并反馈结果

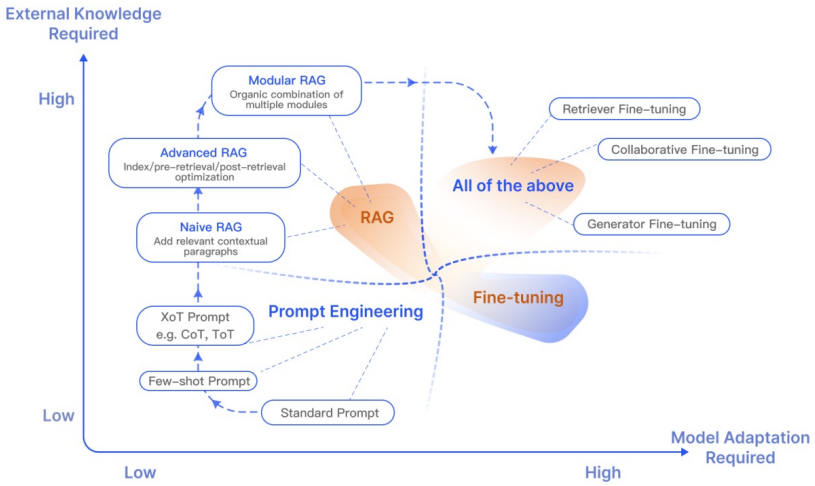
如何提升RAG应用的效果

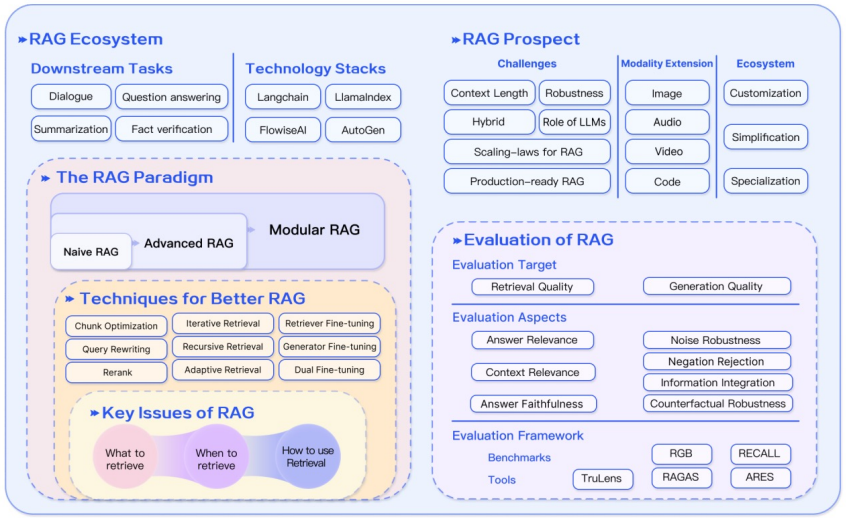
整体效果 = 文档处理效果 * Embedding效果 * Retrieval效果 * LLM效果

Demo容易上生产难



RAG是优先保召回还是优先保精度？





文档处理

- 如何处理原始数据？
- 如何合理地切分Chunk？
- 如何处理不同格式的文档？

Embedding

- 如何选择Embedding模型？
- 如何Fine-tune？
- 运用Embedding最佳实践？

Retrieval

- 如何选择索引和参数？
- 多路召回、Rerank、无效结果处理、MMR
- 如何处理Chunk上下文？

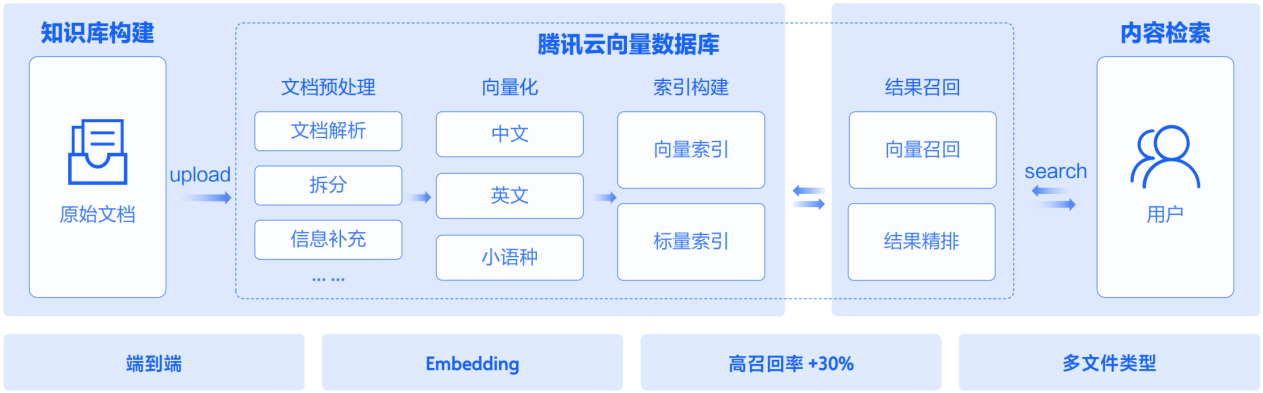
LLM

- Prompts
- Query增强改写
- Query意图识别和路由
- 话题切换
- Fine-tune

腾讯云向量数据库：消除大模型幻觉，加速大模型在企业落地

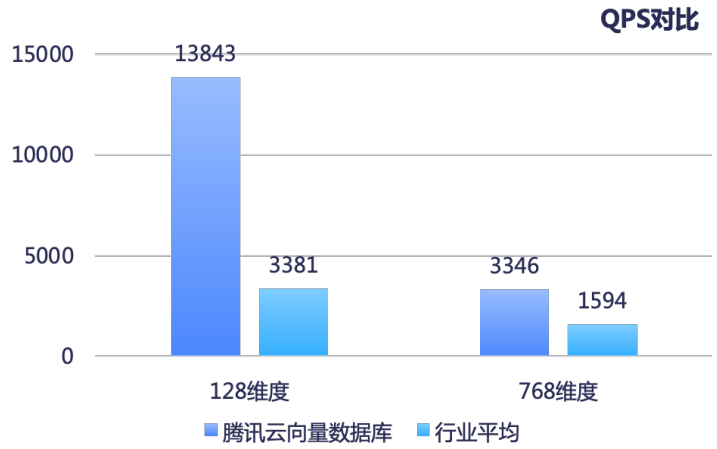
端到端AI套件，AGI时代的知识库解决方案

提供**一站式**知识检索方案，实现业界内**最高召回率、大幅降低开发门槛**，帮助企业快速搭建RAG应用，解决大模型幻觉问题



源自集团多年积累，产品能力行业领先

源自腾讯自研向量检索引擎OLAMA，集团内部**40+**业务线上使用，日均处理**1600亿次**检索请求



首家

『首家』通过中国信通院向量数据库标准测试

千亿

单索引支持最高**千亿级**超大数据规模

500万

单实例最高可达**500万 QPS**