大模型的局限性

1.幻觉问题

面对未知的问题，一本正经的胡说八道。

2.数据更新不及时

我的训练数据在 2023年底截止，之后发生的事情我都无法了解。频繁的更新数据并训练模型耗费资源。

3.缺乏领域知识

通用大模型使用公开数据进行训练，但是缺乏专有领域的知识，解决不了企业专业场景下的问题，而企业训练自己专门的大模型成本又很高。

RAG是一种混合架构，融合两类技术：

检索模型：从大规模知识库（如文档、数据库）中筛选相关信息。

生成模型：基于检索结果和输入问题生成自然语言回答。

与传统模型的区别：

纯生成模型（如GPT）：依赖训练数据中的静态知识，可能生成过时或错误内容。

RAG：实时检索外部知识，生成结果更可靠且可追溯。

RAG技术

通俗解释：

我们可以把 RAG技术（检索增强生成） 想象成一个 “学生开卷考试” 的过程：

1. 考试题目：比如问：“唐朝的科举制度有什么特点？”

   （相当于用户提问的问题）

2. 课本（参考文档）：

   学生手边放着一本《中国历史》教材，里面记录了唐朝的详细历史资料。

   （相当于RAG连接的参考文档）

3. 翻书检索：

   学生先快速翻课本目录，找到“唐朝政治制度”章节，锁定相关内容。

   （RAG的检索阶段：从参考文档中找出与问题最相关的片段）

4. 整合答案：

   学生结合课本里的“科举流程描述”和自己的理解（比如背过的历史意义），组织成一段完整的答案。

   （RAG的生成阶段：大模型将检索到的信息和自己学过的知识融合，生成最终回答）