公开课：《Agentic RAG 原理与实战》

第一部分：原理

1. RAG 背景

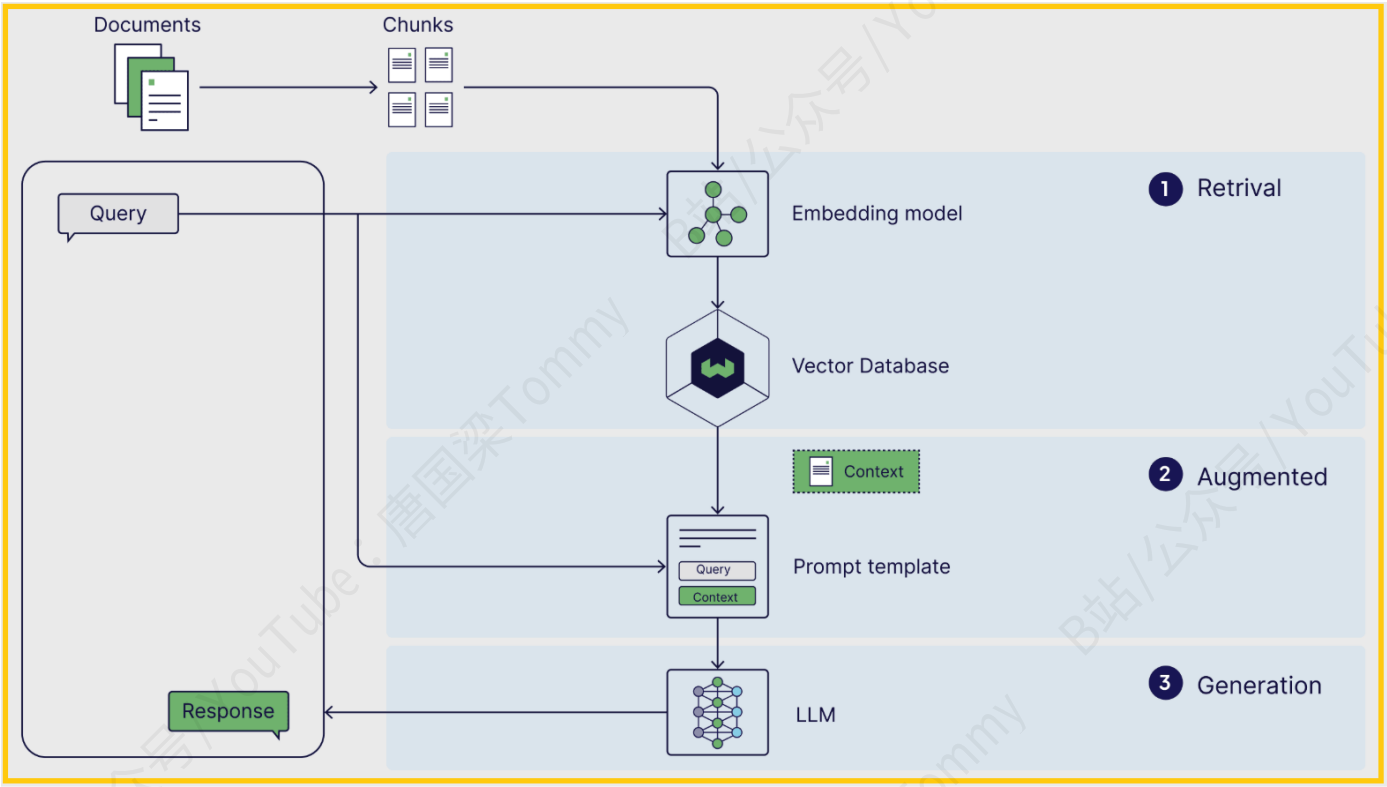
1.1 LLM 面临的挑战

局限	产生原因	典型表现
时效性不足	预训练语料静态，更新不及时	无法回答最近 72 小时新闻、最新法规
行业深度不够	通用语料缺乏垂直领域知识	医疗、法律等专业概念回答不准确
事实幻觉	模型生成依赖语言概率而非事实检索	引用不存在的论文、伪造数据
推理链缺失	黑盒生成，缺乏可解释性与因果链	难以支撑推理过程，合规审核困难

> **核心痛点**：预训练 LLM 本质是「静态知识库 + 语言模式」，一旦问题超出语料覆盖范围，模型极易**凭空编造内容**。

1.2 RAG的作用

RAG（Retrieval-Augmented Generation）通过引入实时检索机制，有效弥补了 LLM 静态知识的局限，增强模型对上下文的理解与时效性响应能力。



RAG 的核心机制包括：

- **Retrieval (R) 检索：** 从外部源、数据库或知识库中搜索相关数据。目标是收集特定、准确和相关的信息，以支持或增强AI对特定主题或查询的理解。
- **Augmentation (A) 增强：** 在此阶段，检索到的数据被添加到提示（prompt）上下文中。这意味着信息被集成或结合到提供给AI的输入中，有效地丰富其知识库，以实现更好的推理和上下文感知响应。
- **Generation (G) 生成：** 最后，LLM 使用增强的上下文生成输出。

借助这一流程，RAG 架构能够显著提升生成内容的**准确性、可控性与现实关联度**，为复杂任务提供更可靠的知识支持。

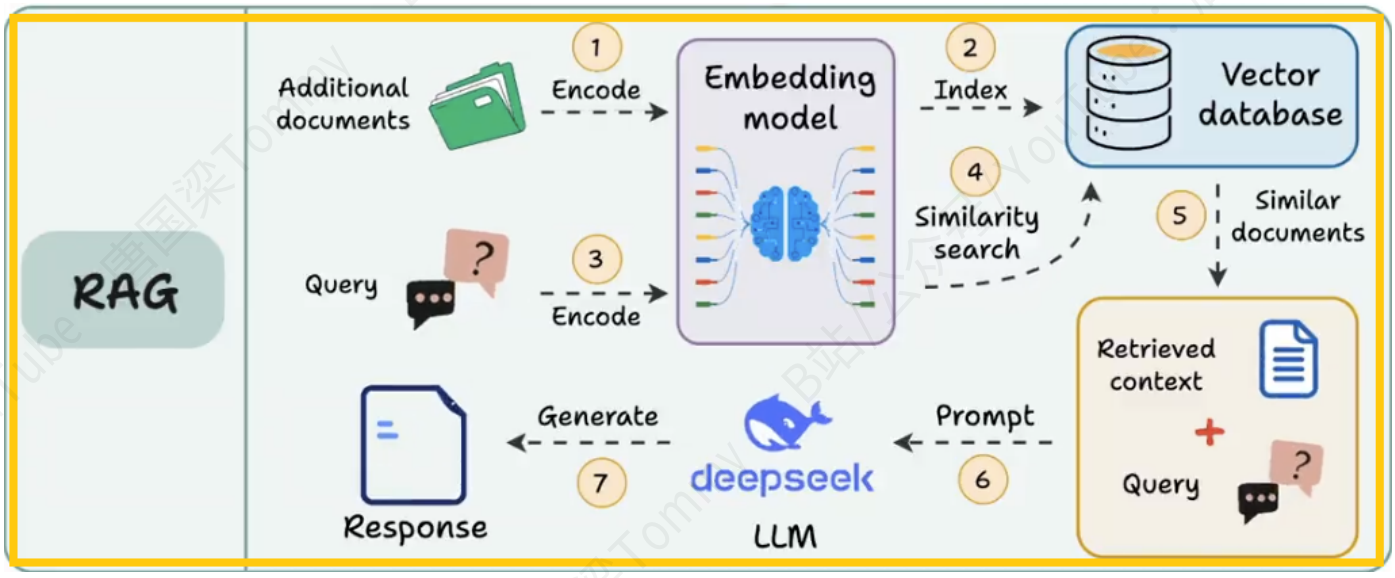
1.3 LLM vs RAG 比较

以下是 RAG和 LLM在多个方面的差异：

特征	LLM（大模型）	LLM With RAG
准确性	容易生成未经证实或幻觉内容，不依赖可靠来源	响应基于外部来源的可验证引用
时效性	依赖静态的预训练数据，可能过时或与当前事件无关	通过整合来自外部源的实时、最新信息增强静态预训练数据
上下文清晰度	经常难以解释歧义查询，导致模糊或不完整的答案	检索特定上下文的信息，提高响应的清晰度
定制化	无法访问或利用用户特定数据集或私有源，导致通用响应	集成公共和私有数据集，实现高度定制化和相关的输出
搜索范围	限于预训练知识库；无法扩展到新信息或外部信息	能够在多个数据库或在线源上进行广泛的、按需搜索
可靠性	由于依赖静态和预生成知识，错误的可能性高	通过实时交叉引用多个受信任的来源确保可靠性
用例	适用于通用任务，但对于动态或数据密集型应用效果较差	适用于需要实时更新、研究或自定义数据集成的任务
透明度	提供的信息没有明确的参考或引用，难以验证	提供引用或链接到来源，确保透明度和可信度

2. RAG 工作原理与挑战

RAG系统架构分为两个主要部分：1️⃣ **数据索引** 2️⃣ **搜索与生成**



2.1 数据索引