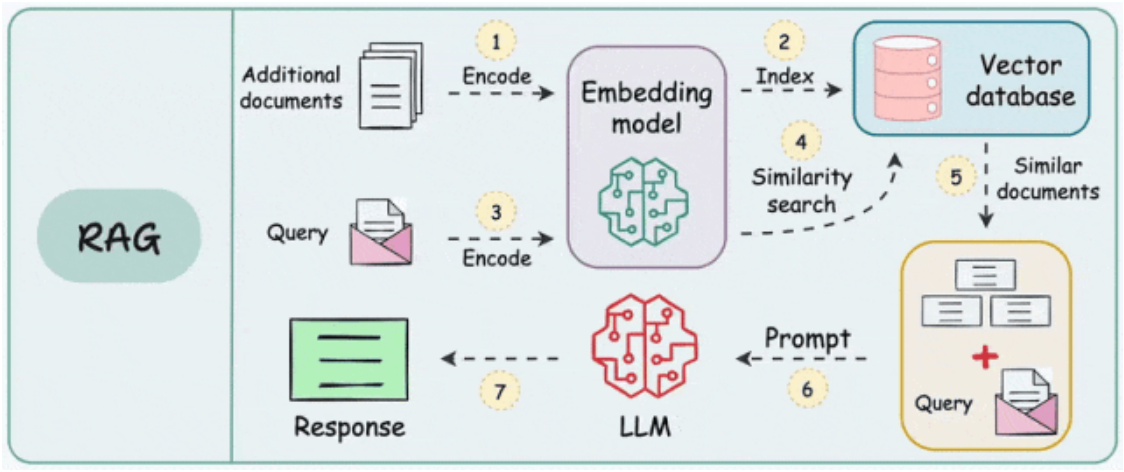
传统 RAG vs. Agentic RAG：动态图示清晰解析

原创 南七无名式 PyTorch研习社 2025年01月18日 11:01 安徽

传统 RAG 存在许多问题：

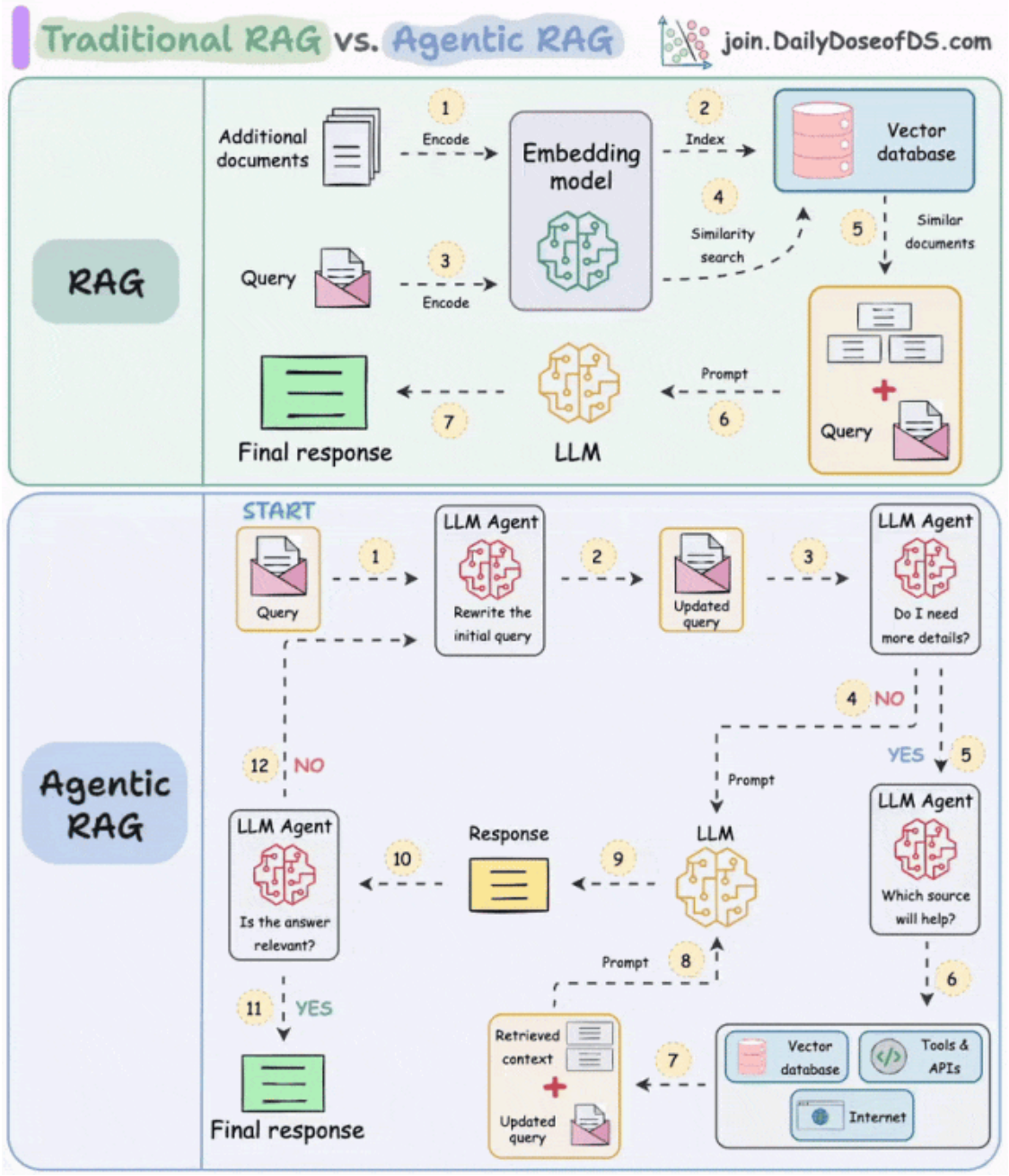
- 它只检索一次并生成一次。如果上下文信息不足，无法动态搜索更多信息。
- 无法处理复杂查询的推理问题。
- 系统无法根据问题调整其策略。



Agentic RAG 尝试解决这些问题。

以下图示展示了它与传统 RAG 的不同之处。

核心思想是在 RAG 的每个阶段引入智能化（Agentic）行为。



第 1-2 步）Agent 会重写查询（如纠正拼写错误等）。

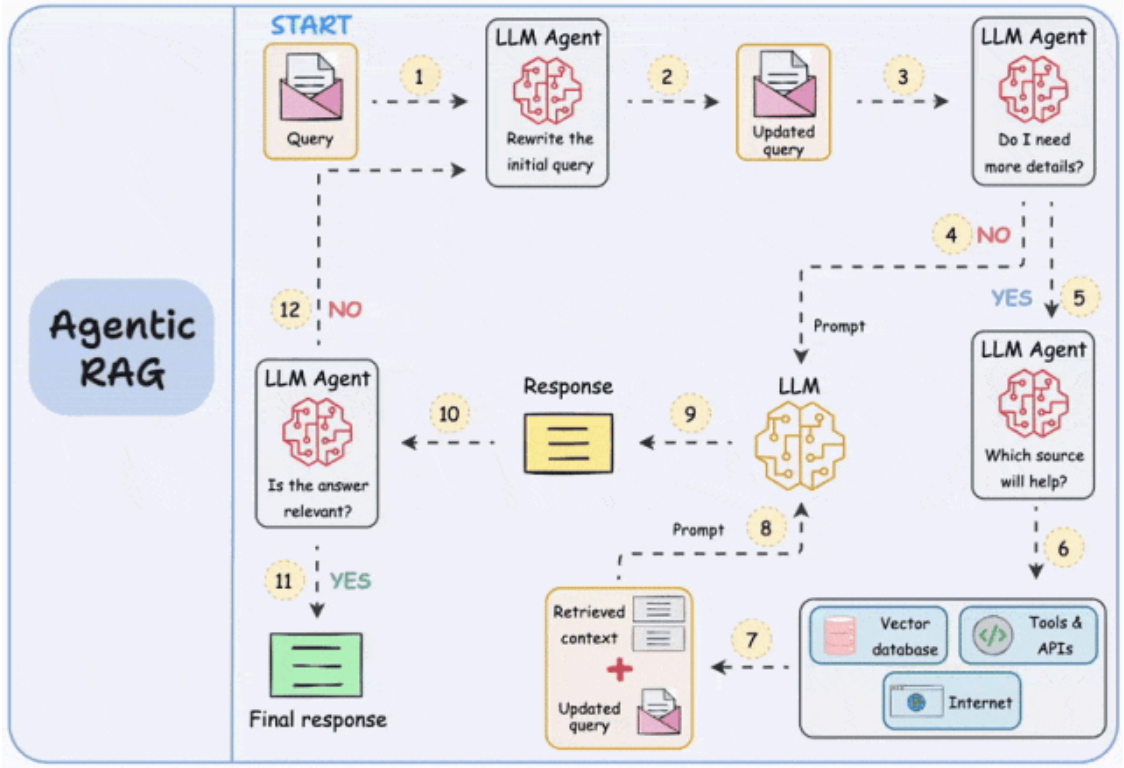
第 3-8 步）Agent 决定是否需要更多上下文信息：

- 如果不需要，重写后的查询直接发送给 LLM。
- 如果需要，智能代理会找到最佳的外部来源以获取上下文，并将其传递给 LLM。

第 9 步）系统生成响应。

第 10-12 步）智能代理检查答案是否相关：

- 如果相关，则返回响应。
- 如果不相关，则返回第 1 步重新开始。



这一过程会重复几次，直到得到合适的回答，或者系统承认无法回答该查询为止。

这使得 RAG 更加健壮，因为 Agent 可以确保每个环节的结果都与目标一致。

需要说明的是，上图只是 Agentic RAG 系统众多架构之一。

你可以根据具体的使用场景对其进行调整和适配。

PyTorch研习社
打破知识壁垒，做一名知识的传播者
648篇原创内容

公众号

RAG 13

RAG · 目录

上一篇
2025年这7种用于构建Agentic RAG系统的架构不可或缺

下一篇
RAG从入门到精通系列3：Routing（路由）