

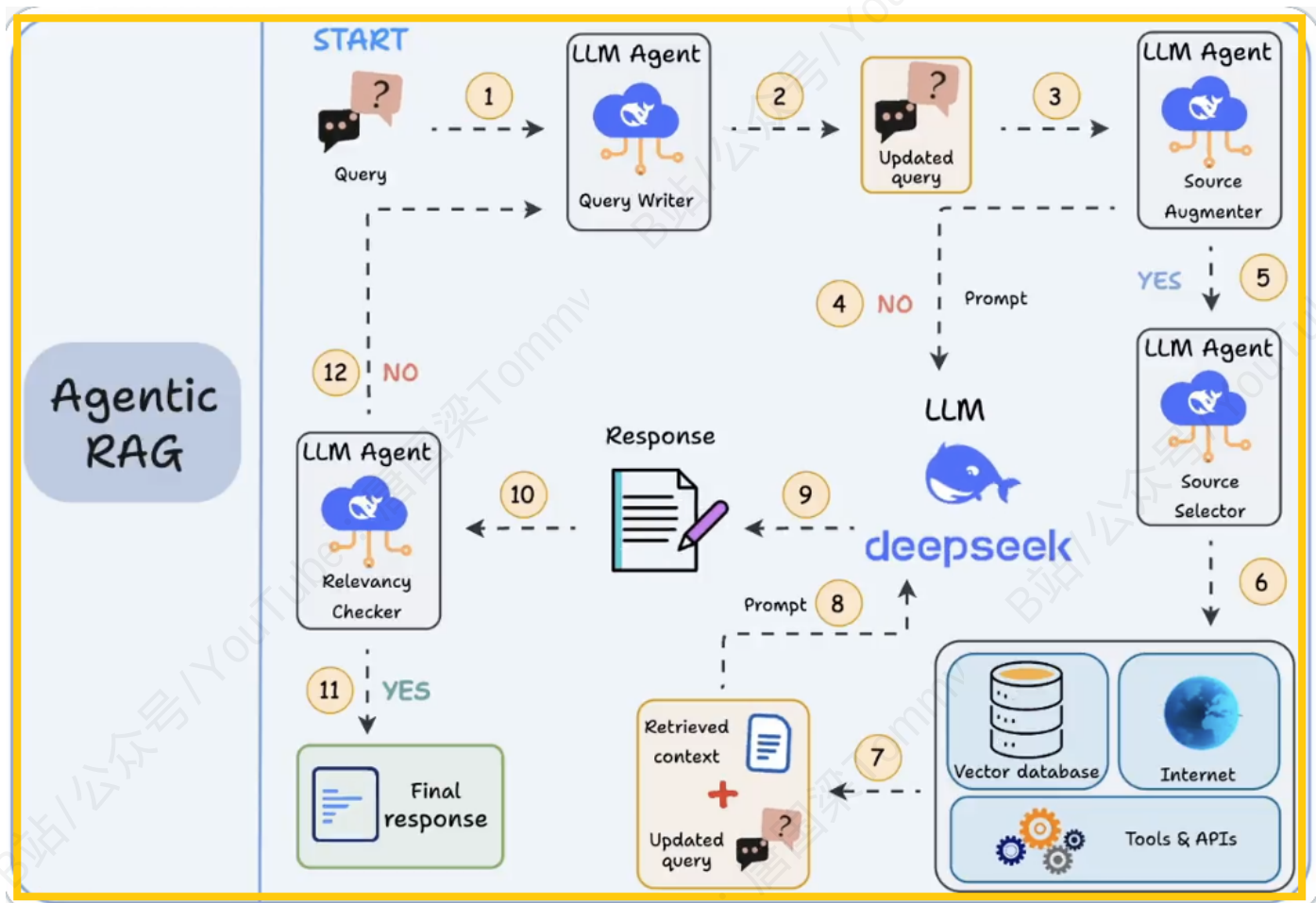
挑战类别	具体挑战	解决方案
1 检索与生成融合	- 检索结果不准确或评分不高导致生成错误 - 多文档融合难，影响上下文利用 - LLM提取答案受上下文噪声影响	- 优化检索算法，提升相关性 - 设计有效的融合策略 - 采用Self-RAG等自反机制提升生成准确性
2 文档切分粒度	- 切块过粗导致语义丢失 - 切块过细带来噪声和上下文不完整	- 基于语义和语言模型辅助分块 - 通过测试调整最佳分块粒度
3 系统扩展性与成本	- API调用成本高 - 向量数据库延迟和吞吐瓶颈 - 维护成本高	- 精简索引，删除冗余 - 分层索引检索减少计算 - 结合本地模型与云端API分配资源
4 隐私与数据安全	- 私有数据传输风险 - 向量数据库存储敏感信息安全问题	- 本地处理或加密存储敏感数据 - 设计访问控制和隐私保护机制
5 检索-生成语义不对称	- 查询与文档语义不匹配 - LLM对矛盾或无关信息鲁棒性差	- 使用假设文档嵌入（HyDE）提升语义对称性 - 利用LLM优化查询 - 引入自然语言推理过滤上下文
7 垂直领域适应性	- 通用模型在专业领域表现差 - 文档拆分和问答分句效果有限	- 领域微调embedding和LLM - 改进拆分和问答分句技术
8 系统有效性验证	- 缺乏统一评估标准 - 需要大量测试和调优	- 建立代表性测试集 - 持续监控和调整模型与检索配置

3. Agentic RAG 概述

Agentic RAG（智能体驱动的检索增强生成）是对传统 RAG 架构的深度演进。通过引入自主 AI 智能体，Agentic RAG 实现了对检索策略的动态管理、上下文理解的迭代优化，以及任务流程的自适应调整，从而突破了传统 RAG 在处理复杂任务和多步推理方面的局限性。

• **核心理念**

- Agentic RAG 的核心在于将智能体引入 RAG 流程，使系统具备以下能力：
- **动态检索策略**：智能体根据任务需求，灵活选择检索工具（如向量数据库、网络搜索、API 接口等），并优化查询方式，以获取最相关的信息。
 - **上下文理解与迭代优化**：通过反思机制，智能体能够评估生成结果的质量，必要时重新检索信息或调整生成策略，提升回答的准确性和相关性。
 - **自适应工作流程**：智能体根据任务的复杂度和类型，动态调整工作流程，实现从顺序执行到多智能体协作的灵活转变。



• 关键特性

- **Agentic (智能体驱动)**：系统具备自主决策和行动能力，能够根据环境变化和任务需求，主动规划和执行操作。
- **RAG (检索增强生成)**：结合外部知识库的信息检索能力与大语言模型的生成能力，提供更准确、上下文相关的响应。

通过融合智能体的自主性与 RAG 的信息整合能力，Agentic RAG 在处理复杂查询、多步推理和动态任务方面展现出更高的灵活性和效率。

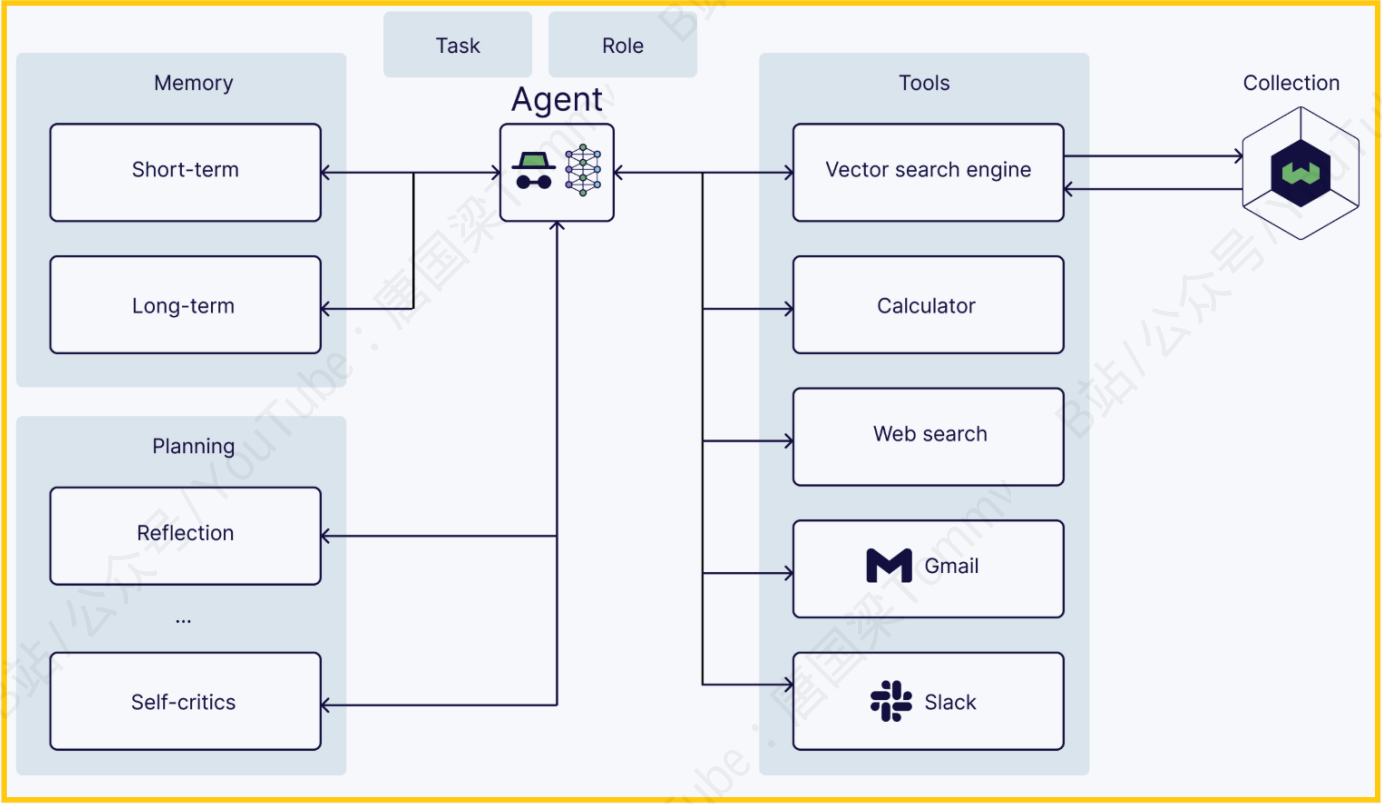
4. Agent 核心组件

Agent智能是 Agentic RAG 系统的基础，使其能够超越传统 RAG 的静态和被动性质。通过集成能够动态决策、迭代推理和协作工作流程的自主智能体(Self Agent)，Agentic RAG 系统表现出增强的适应性和精度。

一个 AI Agent 本质上包含几个组件：

- **LLM (具有定义的角色和任务)**：作为智能体主要的推理引擎和对话接口，负责解释查询、生成响应和维护连贯性。
- **记忆 (短期和长期)**：捕捉跨交互的上下文和相关数据，短期记忆跟踪即时对话状态，长期记忆存储累积的知识和智能体经验。

- **规划（反思和自我批评）**：通过反思、查询路由或自我批评来指导智能体的迭代推理过程，确保复杂任务被有效分解。
- **工具（向量搜索、网络搜索、API 等）**：扩展智能体的能力，使其能够访问外部资源、实时数据或进行专业计算。



5. Agent 设计模式

参考资源

1. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG

<https://arxiv.org/pdf/2501.09136v3>

智能体设计模式为构建具备自主性、适应性和协作能力的智能系统提供了结构化的方法。这些模式使智能体能够在复杂、多变的环境中高效运行，提升系统的准确性和可扩展性。本文重点介绍四种关键模式：

1 反思 (Reflection)

反思模式使智能体能够自我评估并持续优化其输出。通过引入自我反馈机制，智能体可以识别并修正错误、不一致或低效的部分，从而提高整体性能。在实践中，反思通常包括以下步骤：

- **自我评估**：智能体对其输出进行分析，判断其准确性、风格和效率。
- **反馈整合**：将评估结果纳入后续迭代，优化输出质量。
- **外部辅助**：借助单元测试、网络搜索等工具增强评估过程。