**RAG 简介与 RAG 面临的挑战**

检索增强生成（Retrieval Augmented Generation，RAG）是一种连接外部数据源以增强大语言模型（LLM）输出质量的技术。这种技术帮助 LLM 访问私有数据或特定领域的数据，并解决幻觉问题。因此，RAG 已被广泛用于许多通用的生成式 AI（GenAI）应用中，如 AI 聊天机器人和推荐系统。

一个基本的 RAG 通常集成了一个向量数据库和一个 LLM，其中向量数据库存储并检索与用户查询相关的上下文信息，LLM 根据检索到的上下文生成答案。虽然这种方法在大部分情况下效果都很好，但在处理复杂任务时却面临一些挑战，如多跳推理（multi-hop reasoning）或联系不同信息片段全面回答问题。

以这个问题为例：“\_What name was given to the son of the man who defeated the usurper Allectus?\_”

一个基本的 RAG 通常会遵循以下步骤来回答这个问题：

识别那个人：确定谁打败了 Allectus。

研究那个人的儿子：查找有关这个人家庭的信息，特别是他的儿子。

找到名字：确定儿子的名字。

通常第一步就会面临挑战，因为基本的 RAG 根据语义相似性检索文本，而不是基于在数据集中没有明确提及具体细节来回答复杂的查询问题。这种局限性让我们很难找到所需的确切信息。解决方案通常是为常见查询手动创建问答对。但这种解决方案通常十分昂贵甚至不切实际。

为了应对这些挑战，微软研究院引入了 GraphRAG，这是一种全新方法，它通过知识图谱增强 RAG 的检索和生成。在接下来的部分中，我们将解释 GraphRAG 的内部工作原理以及如何使用 Milvus 向量数据库搭建 GraphRAG 应用。

<!--split-->

**GraphRAG 及其工作原理简介**

与使用向量数据库检索语义相似文本的基本 RAG 不同，GraphRAG 通过结合知识图谱（KGs）来增强 RAG。知识图谱是一种数据结构，它根据数据间的关系来存储和联系相关或不相关的数据。

GraphRAG 流程通常包括两个基本过程：索引和查询。

*(图片来源： GraphRAG Paper)*