## 什么是RAG？

RAG是一种把信息检索技术和大模型生成技术结合起来的框架。

## RAG是谁提出来的？

RAG是Facebook的AI研究团队在2020年的NeurIPS（神经信息处理系统会议）上首次提出的概念，具体论文可以查看[Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](https://arxiv.org/pdf/2005.11401v4)

## 什么是KG和LLM？

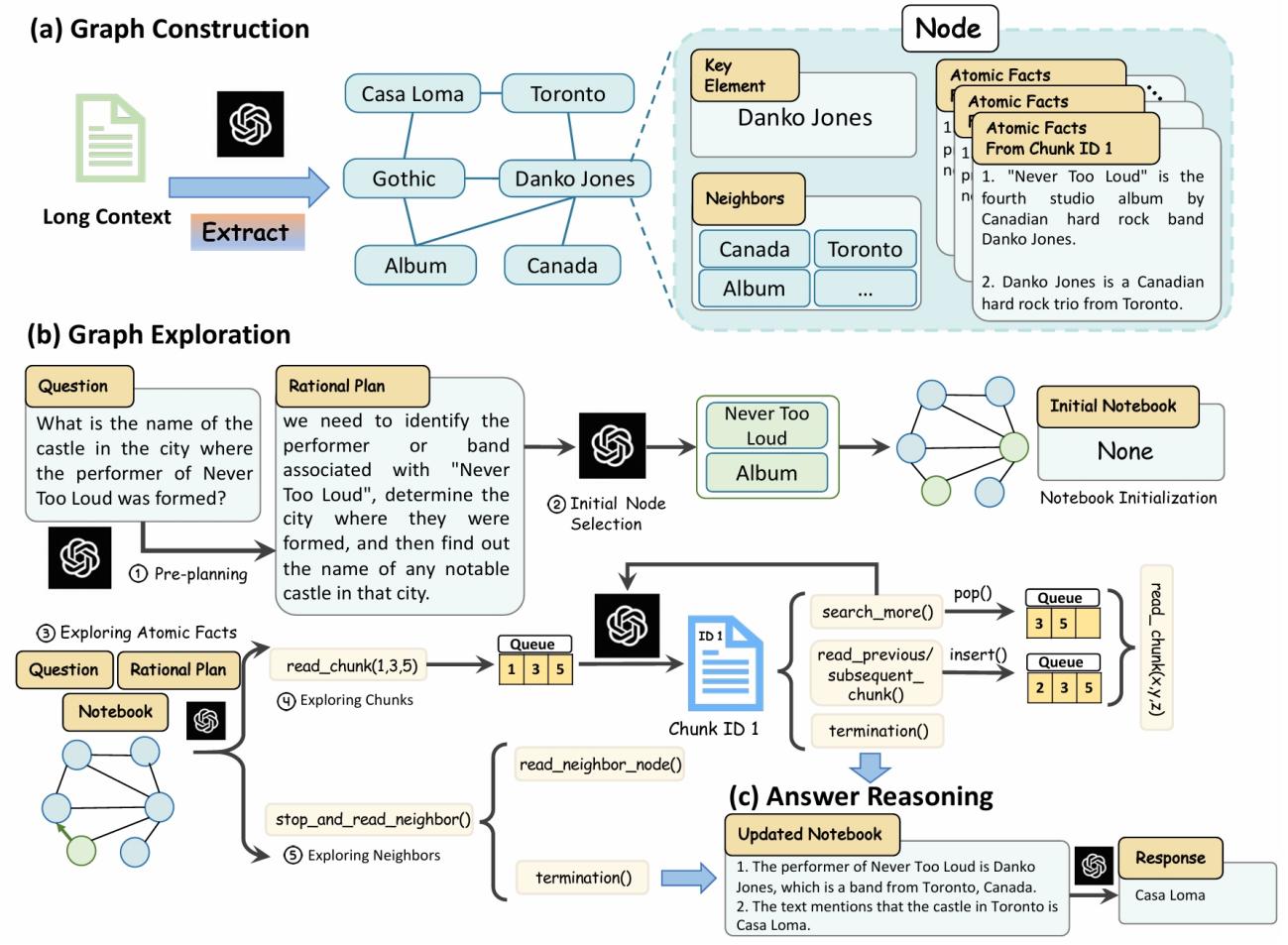
KG是知识图谱的缩写，它在RAG发挥的作用是存储从原始信息提取出的实体关系属性三元组的图数据库。

LLM是大语言模型的缩写，它在RAG发挥的作用是理解自然语言问题和生成自然语言回答。

## RAG的工作原理是什么？

最简单的描述是检索模块找到精准的知识，生成模块优化模型输入输出。

## 用一张图来描述RAG的工作原理。



## 现在的RAG解决了什么问题？

现在的RAG技术已经解决了大模型幻觉问题，能够在生成的回答中标注信息来源。对于医疗救助、法律咨询等对于回答准确性有极高要求的领域有重要帮助。

## 现在的RAG面临什么挑战？

构建多模态数据融合知识库

动态更新知识库数据

压缩模型大小、减少计算时间

保护数据隐私，建立权限体系

引入回答质量评价

## RAG技术的演进历程。

我认为RAG技术从提出到现在只发展了4年时间，提出的技术路线并非越新越好。按照时间顺序排列，这些是已有的RAG技术发展路线。

[Awesome RAG 论文综述](https://github.com/lizhe2004/Awesome-LLM-RAG-Application)

[RAG七十二式：2024年度RAG清单](https://www.cnblogs.com/fanzhidongyzby/p/18599293/rag" \o "发布于 2024-12-11 13:18)

## RAG在政策问答领域的应用。

目前为止，业界最新的应用是牛津大学团队开发的[Medical Graph RAG](https://github.com/SuperMedIntel/Medical-Graph-RAG)，专注于处理医疗数据，已在GitHub开源。

另外有同学对论文[Searching for Best Practices in Retrieval-Augmented Generation](https://arxiv.org/pdf/2407.01219)复现了一篇RAG搭建[最佳实践](https://github.com/chaoql/rag-best-practices)。

同时还有首个专注于政策问答的实践研究：[rag-omni](https://github.com/Logistic98/rag-omni)。