**《RAG极简入门 原理与实践》**

# **检索增强生成模型的探索之旅**

**RAG是什么**

今天我们来聊聊RAG，‌ RAG（Retrieval-Augmented Generation）‌是一种结合了语言模型和 信息检索技术的生成方法，旨在通过检索相关文档来增强语言模型的生成能力，从而提高预测的质量和准确性。

上述概念怎么理解呢？简单的来说就是它让大语言模型（比如ChatGPT或者DeepSeek）在回答你问题时，先去查询相关的资料，再结合自己的资料生成答案，而不是全靠“死记硬背”来回答问题。‌

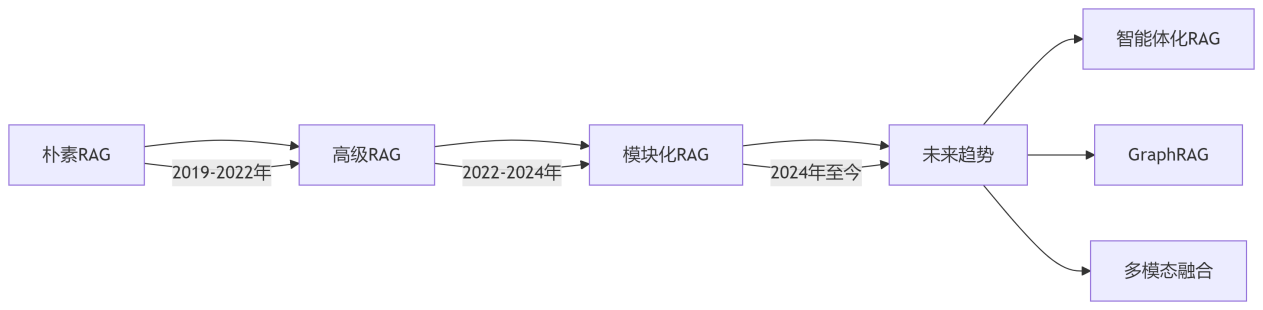
假设你问有一个问题需要LLM来回答：秦始皇统一六国是哪一年？，那么普通的AI和使用RAG方法的AI它们的答案是不一样的，结果精确度上存在明显的差异：  
（1）一般AI（不翻书）：  
全靠自己记忆回答 → 可能答对，容易“瞎编”错误答案（术语叫“幻觉”）

（2）RAG的AI  
翻书：立刻去查《中国历史大事年表》  
找答案：找到“公元前 221 年，秦灭齐，统一六国”

回答问题：根据查到的内容，回答“公元前 221 年”

**RAG发展历程**

RAG（检索增强生成）技术发展历程的总结及对应图示如下：

（1）朴素RAG（Naive RAG，2019-2022年）

索引（Indexing）→ 检索（Retrieval）→ 生成（Generation）的线性流程。

（2）高级RAG（Advanced RAG，2022-2024年）

朴素RAG基础上增加预检索优化、混合检索策略与后检索精炼。

（3）模块化RAG（Modular RAG，2024年至今）

解耦RAG流程为可插拔组件，支持动态编排。

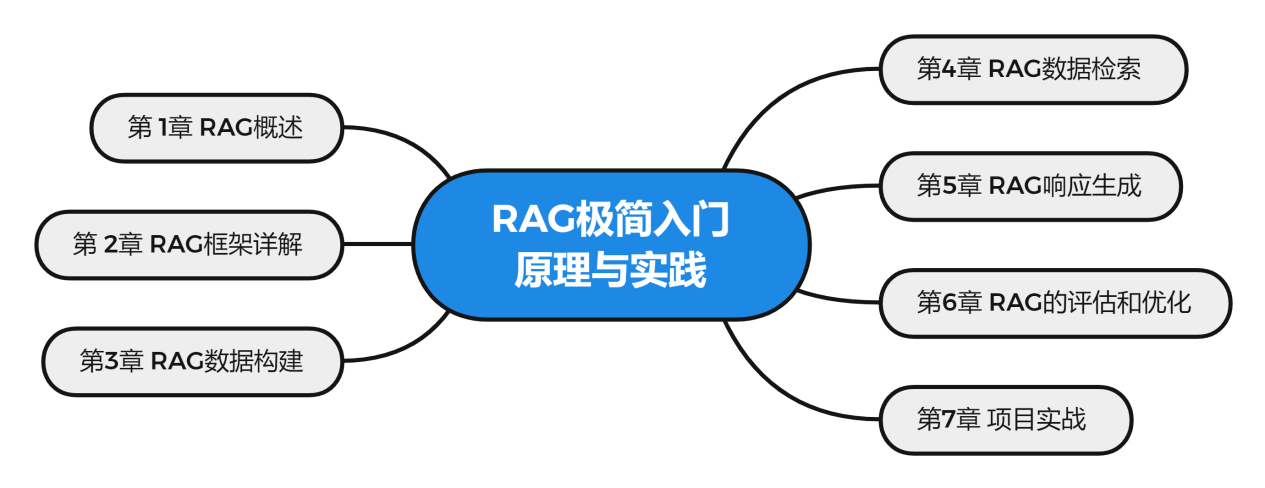
**RAG的极简入门**

我们已经知道，RAG对于LLM是非常重要的，那如何快速了解其原理达到快速入门的效果呢？

这就不得不提及人邮图灵教育的《RAG极简入门原理与实践》这本书了。



本书的目录如下：



内容上还是比较聚焦的，包括RAG概念的阐述，RAG框架的详解，数据构建，数据检索、相应生成、评估和优化以及项目实战。全部围绕RAG的原理和实践来进行。

**为什么需要这本书**

为什么要学习这本书呢？个人觉得至少有以下两点原因：  
**（1）RAG大模型时代的核心基建**

2020年提出RAG框架后，该技术因能有效解决大模型“幻觉”问题，被Google、微软等巨头广泛应用于知识管理、企业智能客服等领域。但是呢随着企业数据量激增，单纯依赖长上下文窗口模型面临显存消耗平方级增长（如128K上下文首Token延迟达数十秒）与成本过高的瓶颈，最终使得RAG“检索-压缩-生成”的扩展方式成为必选项。

**（2）系统化指南**

现有学习RAG的资料分散且偏理论，很多人陷入“检索器选型→向量化优化→生成控制”的细节迷局。本书以“极简”为切口，将企业级RAG系统拆解为7大模块（全书的7大章节）、47项关键技术点（如触发策略设计、知识冲突解决），形成了从理论到落地的完整系统化指南。

**本书特点**☑ 全面且深入浅出

本书从 RAG 的基本概念入手，逐步深入到框架详解、数据构建、数据检索、响应生成、评估优化以及项目实战等各个环节，构建了一个完整且逻辑清晰的知识体系。同时，对 RAG 和长上下文的讨论，也为后续深入理解 RAG 的应用场景奠定了基础。

后文中对 RAG 框架中触发检索的判断策略、大模型自身知识和检索知识平衡的细致剖析，还是对 RAG 数据构建过程中向量化技术、向量数据库、数据解析与处理的全面讲解都非常的细致。体现作者水平的强大。

☑ 理论和时实践想结合

本书的一大亮点在于将理论知识与实践操作完美融合，如：

1. 第3章 “RAG 数据构建” 中，详细讲解了向量化技术在 RAG 中的作用以及向量模型的评估与选择，随后通过基于 DeepSeek 和 Ollama 的代码实践，让读者能够亲自动手操作，加深对理论知识的理解。



上述代码简单展示了用户查询到回答结果的完整流程，流程清晰注释完全。

1. 第7章“项目实战”中，更是涵盖了从搭建基础 RAG 系统到优化检索模块、增强生成模块、与知识图谱结合实践、多模态 RAG 实现，以及系统优化与调试、构建端到端应用。

这种理论与实践相结合的写法，使得本书既具有学术性，又具有很强的可实践的操作性，能够满足不同读者的学习需求。

☑ 内容新颖且前沿

目前RAG 领域不断涌现出新的研究成果和应用实践。

本书紧跟技术前沿，在介绍RAG框架时，涵盖了常见的RAG范式：



还对一些新兴的RAG技术进行了探讨：

1. 第5章“RAG 响应生成”中提到的大模型选择与优化、解码策略、融合外部知识增强 RAG 生成、多模态扩展以及主动问答与交互能力等内容。
2. 第6章“RAG 的评估和优化”中，对动态相关 RAG、Graph RAG、FlashRAG、DocReLM 等前沿 RAG 方法的介绍，更是让读者能够及时了解到 RAG 领域的最新发展动态，拓宽了读者的视野，激发了读者对RAG技术未来发展的思考。

**多元化的受众群体**

本书的受众群体也是比较丰富的：

☑ AI与NLP领域的学生和研究人员

对于正在学习人工智能和自然语言处理的学生来说，本书是一本难得的教材。本书不仅能够帮助他们快速了解 RAG 技术的最新研究进展和应用成果，还能够为他们提供一些新的研究思路和方法。

☑ 软件工程师和数据科学家

软件工程师可以了解如何将RAG技术应用到实际的软件开发项目中，例如开发智能问答系统、智能写作助手、知识图谱应用等。书中提供的项目实战案例和代码，能够帮助他们快速掌握 RAG系统的搭建和优化方法，提升软件系统的智能化水平和用户体验。

☑ 对AI感兴趣的行业从业者和爱好者

书中通俗易懂的语言和丰富的实例，对于AI爱好者来说这些实例可以帮助读者快速掌握RAG的原理和核心,即使是没有深厚技术背景的读者也能够轻松理解RAG技术的实战方法。