构建企业级大模型应用：RAG 技术实战与应用

讲师 林老师

【课程介绍】

这是一门针对大模型工程开发和应用人员的课程。全部课程大约 21 课时。

“当大模型在专业领域一本正经地胡说八道时，企业如何避免千万级 GPU 算力打水漂？”

面对大模型固有的知识滞后性和行业数据安全要求，企业亟需一种低成本、零训练、实时更新的解决方案——这正是 RAG（检索增强生成）技术的爆发点。

与传统的关键词搜索相比，RAG 系统因为大模型能够理解语义、总结语义、生成自然语言回答，在提升信息获取效率、增强知识利用率的方面有着无以伦比的优势。比起训练一个大模型，RAG 系统建设的投入成本要小得多。因此，RAG 成为大模型应用快速落地的关键技术。

本课程直击大模型应用三大痛点：

破除幻觉：教会模型“知之为知之”的检索式对话能力

解放算力：免训练接入私有数据

动态更新：10 分钟完成企业知识库实时热更新

本课程深入解析 RAG（Retrieval-Augmented Generation）技术如何解决大模型知识滞后与幻觉问题，通过全流程实战教学带学员使用 Python 快速构建企业级 RAG 系统。课程覆盖数据预处理→向量检索→排序优化→大模型生成完整技术链，基于 deepseek、qwen 等主流模型，结合 LangChain、LlamaIndex、FAISS 等工具在 Windows+Anaconda 环境下实现可落地的 RAG 解决方案。

【课程收益】

1、直击 RAG 原理： 掌握 RAG 核心技术。

2、掌握构建 RAG 核心技术： 掌握通用 RAG 技术知识，针对特定场景下，对 RAG 流程优化，使其更贴合实际应用。

3、丰富项目实战经验： 通过丰富的场景和项目实战，所学即所用，积累宝贵的实战经验。

【课程特色】

1、案例驱动，注重实操： 课程以案例方式进行知识内容的梳理和学习，配合实操环节学员都能掌握项目实操。

2、理论与实践结合： 既深入讲解理论技术、配合技术实操，帮您快速了解原理，同时应用到时间工作中。

3、一线专家亲授： 由一线的实践经验丰富的专家授课，分享实战经验。

【课程大纲】

1、基本常识概述（1 小时）

1）人工智能概况

2）RAG 起源与趋势

3）RAG 基本原理与价值

4）RAG 常见问题

2、大模型原理剖析与应用（4 小时）

1）大语言模型基本原理

2）大模型的应用场景

3）提示词工程

4) 大模型部署简介和应用开发环境准备

5）大模型应用（RAG、智能体）开发

3、快速搭建自己的 RAG 系统（5 小时）

1）用 LIama Index 快速搭建 RAG 系统

2）用 Lnagchain 快速搭建 RAG 系统

3）用 python 快速搭建 RAG 系统

4) 不同参数对 RAG 效果影响分析

4、企业级 RAG 优化技术（6 小时）

1）常见文档切分与处理

2）常见的向量化模型

3）常见的 RAG 优化技巧

4）纠正检索增强生成（Corrective-RAG）

5）检索增强微调（RAFT）

6）智能体检索增强（Agentic RAG）

7）企业知识库检索项目实战

6、热门开源 RAG 相关项目应用（5 小时）

1）RAGFlow 应用

2）QAnything 应用

3）Dify 开发框架

【环境准备】

1. 本课程提供项目实战代码。运行代码的硬件环境如下：
2. Intel i5 以上处理器，32G 以上内存，100G 以上硬盘空间；
3. 操作系统为 windows；
4. 安装 python 环境 3.10+，推荐使用 Anaconda。

【课程对象】

- AI 工程师/算法开发者
- NLP 技术爱好者（需基础 Python 知识）
- 企业技术团队负责人（了解技术边界）
- 数据科学家（需升级大模型应用能力）
- 技术型产品经理（掌握 RAG 技术实现逻辑）

【授课方式】

本课程采用直播授课模式。

【讲师简介】

林老师，人工智能领域高级工程师。拥有近 15 年 AI 领域实战经验，主导搭建包含 Qwen、DeepSeek 等 10 余款大语言模型企业级推理服务平台，并主导构建覆盖研发、制造、市场等多领域的智能知识库系统。具备基于私域知识微调大模型的经验，深度参与从传统知识库到智能问答引擎的技术跃迁，成功主导/交付多个企业级知识库架构规划及项目落地。