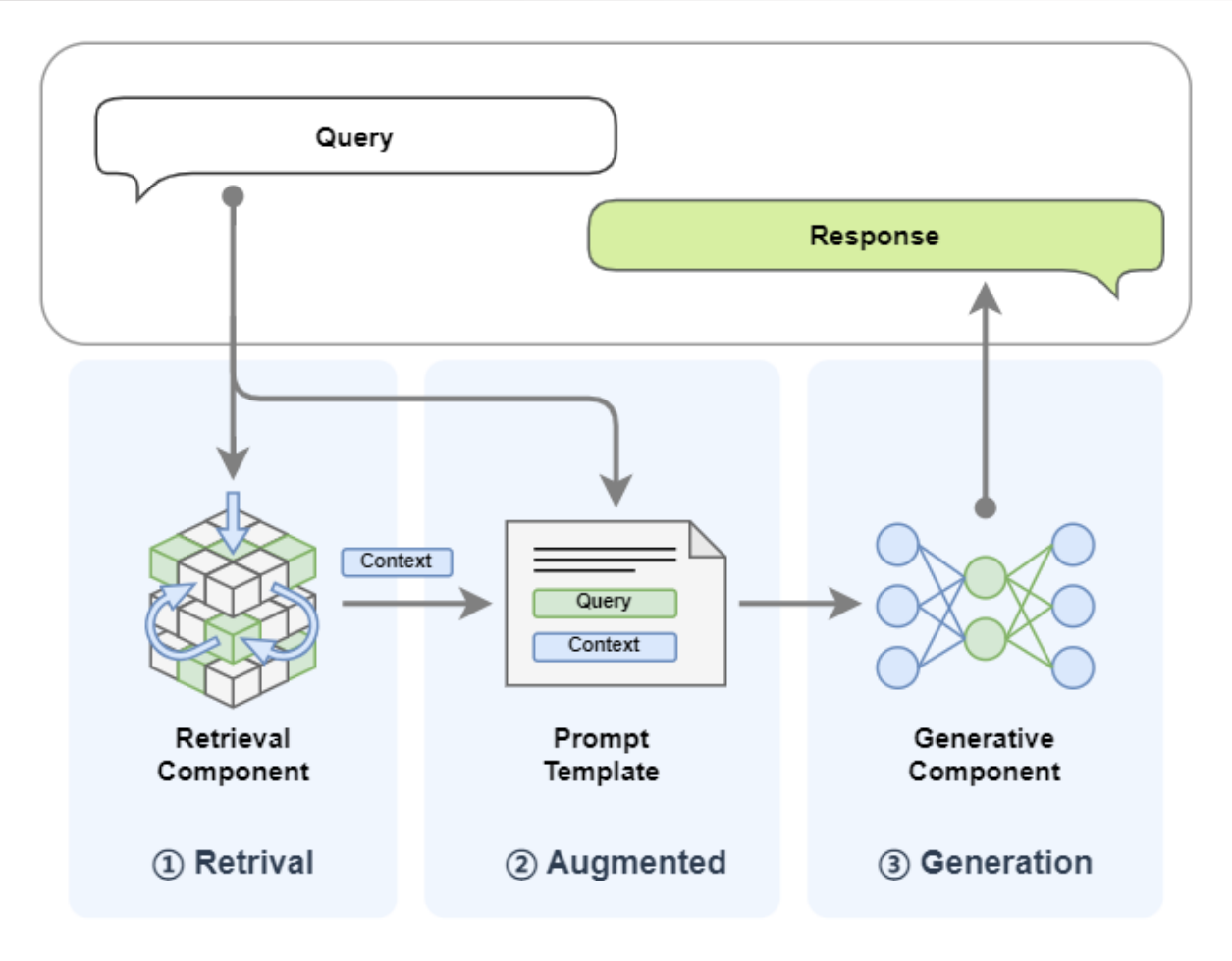
RAG 技术详解与实践应用

第0讲：课程介绍

目录

1. 课程简介
2. 讲师介绍
3. 选题初衷
4. 课程导览

课程简介



以RAG(Retrieval-Augmented Generation)为核心，
从基础的检索与生成策略入手，
深入探索效果优化与性能提升，
逐步拓展多模态能力，
强大的智能Agent全面赋能，
构建可落地的企业级知识库。
通过大量的实战案例，
专业的导师手把手教学，
获得从原型开发到系统工程化的完整能力成长，
助力您斩获心怡的offer！

目录

1. 课程简介
2. 讲师介绍
3. 选题初衷
4. 课程导览

讲师介绍

王志宏 商汤科技大装置事业群研发总监

- 商汤的私有化AI综合解决方案研发负责人，并主导开源项目 LazyLLM 的技术研发与生态建设
- 前商汤自研的深度学习训练框架SenseParrots的研发负责人
- 荣获开源中国源创会2024年度技术领航者称号
- 深耕 AI 领域多年，具备丰富的企业级RAG 实践经验，推动数十家企业实现 AI 应用落地，累计为企业创造了数千万元的商业价值

目录

1. 课程简介
2. 讲师介绍
3. 选题初衷
4. 课程导览

大模型时代的技术选型

技术方向	主要技能	学习难度	算力门槛	实用性
预训练 Pre-training	深度学习基础：反向传播、优化算法、梯度下降 自然语言处理：词向量、Transformer、Attention 机制 训练框架：PyTorch 大规模分布式训练：分布式计算、各种并行计算	★★★★	★★★★ ★★	★
微调 Fine-tuning	监督学习：分类、生成、问答任务建模和数据清洗 模型训练与优化：调参，过拟合处理 训练框架：PyTorch	★★	★★	★★
RAG	文本处理：文本读取，解析和切分策略 向量化检索：词向量化，向量数据库 模型使用：大模型，多模态模型，... 应用部署：服务化、API 接口开发，界面开发	★	★	★★★ ★★
智能体 Agent	任务规划与调度：多任务管理、任务拆解 环境交互：API 集成、插件开发	★★	★	★★★
强化学习 RL	强化学习：学会一个强化学习的训练框架 算法定制：制定目标、Reward模型和Policy模型	★★★★ ★	★★★★	★★

RAG 的应用落地性最强，学习成本相对可控，因此是当前**最实用**、**最具商业价值**的方案。

大模型时代的技术选型

	从 0 到 1 手搓 RAG	基于已有框架搭建并逐步优化效果
方案成熟度	对大部分人而言，稳定性和性能都不太好	稳定性和性能都比较好
开发周期	开发周期长，工作量大	快速搭建 MVP，进行效果验证
学习门槛	对代码能力要求较高	完成基础目标所需要的代码能力较低
定制化程度	灵活性高，可任意定制	不同框架的灵活度不同，但定制化程度均受限
学习重心	深入理解 RAG 工作原理	可快速聚焦优化检索效果和生成质量
适用用户群体	1. 想锻炼代码能力的人 2. 觉得已有框架不能满足需求的群体	1. RAG初学者 2. 企业开发者

结论：对于初学者而言，**基于已有框架**是明智的选择：

1.快速落地，先见成效：已有框架提供了完备的检索、生成、调用接口，大大缩短开发周期。
2.成熟度高，性能优化好：框架解决了很多复杂的工程问题，避免重复“造轮子”。
3.聚焦业务，注重成效：专注应用场景的业务逻辑优化，而非底层细节，聚焦关键目标。

目录

1. 课程简介
2. 讲师介绍
3. 选题初衷
4. 课程导览

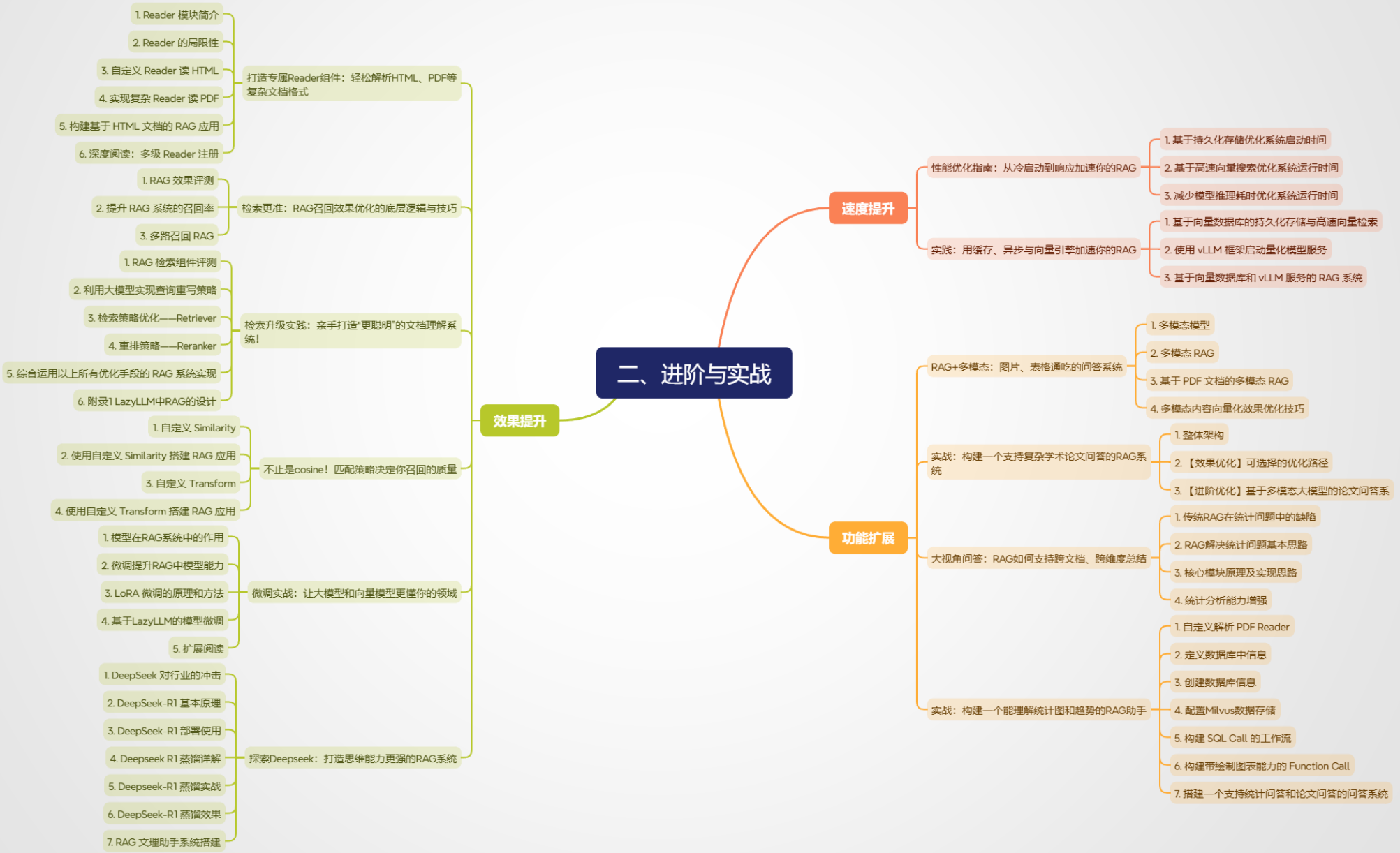
课程导览


LLM与RAG技术：原理、应用与实战
一、基础与实战
RAG原理解读：让检索增强生成不再是黑盒
10分钟上手一个最小可用RAG系统
大模型怎么玩？用LazyLLM带你理解调用逻辑与Prompt魔法
RAG项目工程化入门：从脚本走向模块化与可维护性
二、进阶与实战
效果提升
打造专属Reader组件：轻松解析HTML、PDF等复杂文档格式
检索更准：RAG召回效果优化的底层逻辑与技巧
检索升级实践：亲手打造"更聪明"的文档理解系统！
不止是cosine！匹配策略决定你召回的质量
微调实战：让大模型和向量模型更懂你的领域
探索Deepseek：打造思维能力更强的RAG系统
速度提升
性能优化指南：从冷启动到响应加速你的RAG
实践：用缓存、异步与向量引擎加速你的RAG
功能扩展
RAG + 多模态：图片、表格通吃的问答系统
实战：构建一个支持复杂学术论文问答的RAG系统
大视角问答：RAG如何支持跨文档、跨维度总结
实战：构建一个能理解统计图和趋势的RAG助手
三、企业级RAG
企业级RAG：权限、标签、拒答机制与内容安全的全链路方案
四、高阶RAG与实战
RAG x Agent：构建可调用工具、可自我反思的智能助手
RAG x 知识图谱：从关系结构中召回更准确的内容

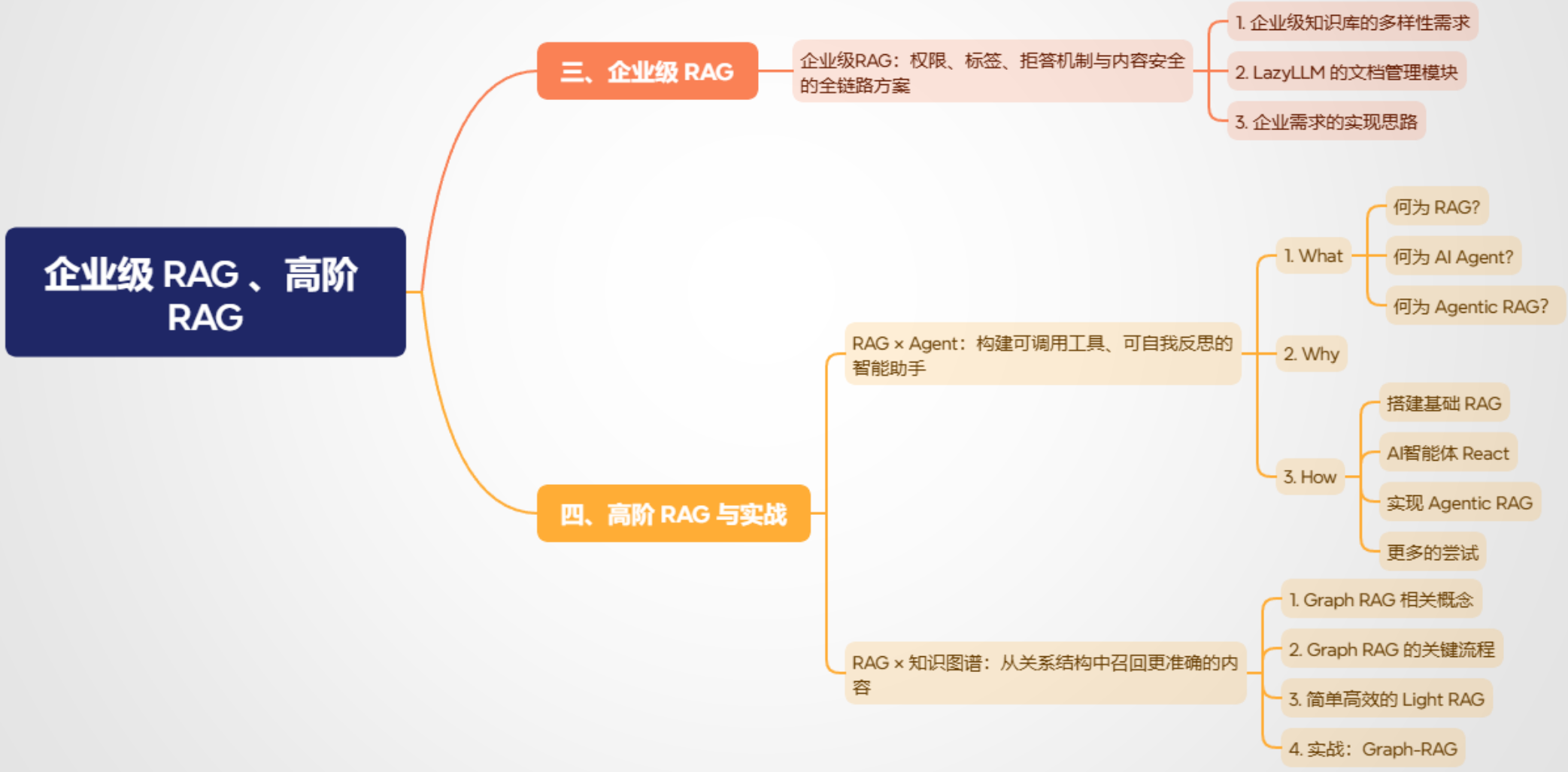
课程导览


一、基础与实战
RAG 原理解读：让检索增强生成不再是黑盒
1. 大模型的基本概念
2. 从大模型到 RAG
3. RAG 的基本概念
10分钟上手一个最小可用 RAG 系统
1. RAG 实践剖析
2. RAG 知识库构建
3. 实现基础的 RAG
大模型怎么玩？用 LazyLLM 带你理解调用逻辑与 Prompt 魔法
1. 数据流简介
Pipeline、Parallel、Diverter等
2. 调用大模型
在线大模型
本地大模型
RAG 项目工程化入门：从脚本走向模块化与可维护性
1. 代码管理：从零开始搭建 Git 项目
2. 版本管理和分支策略
3. 代码规范与自动检查
4. 单元测试与代码质量保障
5. 持续集成（CI）：让测试自动化
6. 代码文档管理：用 MkDocs 搭建文档系统
7. 制品

课程导览


二、进阶与实战
效果提升
打造专属Reader组件：轻松解析HTML、PDF等复杂文档格式
1. Reader 模块简介
2. Reader 的局限性
3. 自定义 Reader 读 HTML
4. 实现复杂 Reader 读 PDF
5. 构建基于 HTML 文档的 RAG 应用
6. 深度阅读：多级 Reader 注册
检索更准：RAG召回效果优化的底层逻辑与技巧
1. RAG 效果评测
2. 提升 RAG 系统的召回率
3. 多路召回 RAG
检索升级实践：亲手打造"更聪明"的文档理解系统！
1. RAG 检索组件评测
2. 利用大模型实现查询重写策略
3. 检索策略优化——Retriever
4. 重排策略——Reranker
5. 综合运用以上所有优化手段的 RAG 系统实现
6. 附录1 LazyLLM中RAG的设计
不止是cosine！匹配策略决定你召回的质量
1. 自定义 Similarity
2. 使用自定义 Similarity 搭建 RAG 应用
3. 自定义 Transform
4. 使用自定义 Transform 搭建 RAG 应用
微调实战：让大模型和向量模型更懂你的领域
1. 模型在RAG系统中的作用
2. 微调提升RAG中模型能力
3. LoRA 微调的原理和方法
4. 基于LazyLLM的模型微调
5. 扩展阅读
探索Deepseek：打造思维能力更强的RAG系统
1. DeepSeek 对行业的冲击
2. DeepSeek-R1 基本原理
3. DeepSeek-R1 部署使用
4. Deepseek R1 蒸馏详解
5. Deepseek-R1 蒸馏实战
6. DeepSeek-R1 蒸馏效果
7. RAG 文理助手系统搭建
速度提升
性能优化指南：从冷启动到响应加速你的RAG
1. 基于持久化存储优化系统启动时间
2. 基于高速向量搜索优化系统运行时间
3. 减少模型推理耗时优化系统运行时间
实践：用缓存、异步与向量引擎加速你的RAG
1. 基于向量数据库的持久化存储与高速向量检索
2. 使用 vLLM 框架启动量化模型服务
3. 基于向量数据库和 vLLM 服务的 RAG 系统
功能扩展
RAG+多模态：图片、表格通吃的问答系统
1. 多模态模型
2. 多模态 RAG
3. 基于 PDF 文档的多模态 RAG
4. 多模态内容向量化效果优化技巧
实战：构建一个支持复杂学术论文问答的RAG系统
1. 整体架构
2. 【效果优化】可选择的优化路径
3. 【进阶优化】基于多模态大模型的论文问答系
大视角问答：RAG如何支持跨文档、跨维度总结
1. 传统RAG在统计问题中的缺陷
2. RAG解决统计问题基本思路
3. 核心模块原理及实现思路
4. 统计分析能力增强
实战：构建一个能理解统计图和趋势的RAG助手
1. 自定义解析 PDF Reader
2. 定义数据库中信息
3. 创建数据库信息
4. 配置Milvus数据存储
5. 构建 SQL Call 的工作流
6. 构建带绘制图表能力的 Function Call
7. 搭建一个支持统计问答和论文问答的问答系统

课程导览


企业级 RAG 、高阶 RAG
三、企业级 RAG
企业级RAG：权限、标签、拒答机制与内容安全的全链路方案
1. 企业级知识库的多样性需求
2. LazyLLM 的文档管理模块
3. 企业需求的实现思路
四、高阶 RAG 与实战
RAG × Agent：构建可调用工具、可自我反思的智能助手
1. What
何为 RAG?
何为 AI Agent?
何为 Agentic RAG?
2. Why
3. How
搭建基础 RAG
AI智能体 React
实现 Agentic RAG
更多的尝试
RAG × 知识图谱：从关系结构中召回更准确的内容
1. Graph RAG 相关概念
2. Graph RAG 的关键流程
3. 简单高效的 Light RAG
4. 实战：Graph-RAG

Q&A

1. 为什么预训练的实用性只有一颗星
2. 课程上线节奏是什么样的

感谢聆听

Thanks for Listening