- RAG
  - RAG三大组件
    - 大语言模型
      - 基础知识
        - 文字接龙，预测下一个token，串行执行
        - token
        - transformer/self-attention
      - 训练过程
        - 预训练
        - 指令微调
        - 强化学习:RLHF
      - 产品：国内外主流大模型产品
      - 大模型产品选型
        - 大模型涌现能力
        - 模型评测过程
      - 为RAG挑选大模型
        - 模型大小
        - 能力评测
        - 成本和设备
        - 数据安全
      - 使用大模型
        - GPU知识
        - transformers
        - ollama
        - 第三方产商API
        - 流式输出
    - 知识库
      - embedding模型
        - 架构：基于transformer encoder的BERT
          - 完型填空
          - 预测下一句
        - embedding基准评测：MTEB
        - 两大主流中文embedding模型：GTE/BGE
          - 相似性度量学习
          - 对比学习
        - 如何挑选embedding模型：MTEB+迭代
        - 使用embedding模型
          - 3种调用方式
          - 相似度计算
          - 聚类计算
      - 向量数据库
        - 产品介绍和全方位对比
          - chroma
          - milvus
        - 索引技术
          - FLAT Index
          - IVF
          - PQ
          - HNSW
          - 建立合适的索引建议
        - 使用chroma/milvus
          - 连接
          - 元数据
          - 构建集合
          - 索引
          - 查询
      - 文件解析和分块技术
        - 复杂性：企业数据的复杂多样
        - 质量为王：GIGO原则
        - 文件解析
          - pdf
          - html
          - txt/markdown/json/xml
          - word/ppt/excel
          - RAGFlow：deep-doc
        - 分块技术
          - 递归文本分块
          - 基于embedding的语义分块
          - 基于模型的语义分块
      - Graph图数据
        - 三元组：基于实体和关系构建的图关联的知识组织，本身具备强关联性
        - 两大产品
          - nebulagraph
          - neo4j
            - 安装
            - cypher语句
              - 创建节点
              - 创建关系
              - MATCH查询
            - py2neo
    - 检索
      - 查询增强
        - Query2Doc
        - HyDE
        - 子问题查询
        - 查询改写
        - Take a Step Back
      - 多索引增强
        - 父子索引
        - 总结索引
        - 假设问题索引
      - 检索后增强
        - Rerank重排
        - 混合索引
      - 模块检索
        - 迭代检索增强生成
        - 新范式：self-RAG
  - RAG评估
    - 评估标准
      - 生成答案和上下文：忠实性
      - 生成答案和问题：答案相关性
      - 上下文和问题：上下文相关性
    - RAGAS
      - 评估数据构建
      - 评估指标
        - 忠实性
        - 答案相关性
        - 上下文召回
        - 上下文精度
      - 执行评估
    - 评估后：有针对性的迭代优化
  - RAG+
    - RAG+GRAPH
      - GraphRAG特点
        - 多跳关系查询：在xxx的同学的朋友中，谁在阿里巴巴工作?
        - 语义关联查询：还有哪些采用超感光摄影的手机?
        - 综合性查询：最近几年高端智能手机的整体发展趋势是怎样的?
      - 构建GraphRAG
        - 提取问题中的关键词信息
        - 根据关键词信息检索图数据库
        - 拼接检索得到信息作为上下文，送到大语言模型来生成答案
    - RAG+Agent
      - ReAct Agent
        - Reason
        - Action
        - 用Reason来指导Action，同时Action的Observation来辅助Reason
        - 模式：Thought-Action-Observation
      - RAG-router
    - RAG+微调
      - LLM微调
        - 什么时候需要微调
        - PEFT：参数高效的微调 LoRA
        - 大语言模型微调框架swift
      - embedding模型微调
  - RAG项目
    - 制度问答助手
    - 基于知识图谱的金融智库问答助手
    - gradio界面
  - AI应用开发软技能
    - LLM需求能力分析：不同项目角色需要对AI大模型的了解程度差异性分析
      - 产品经理
      - AI应用开发人员
      - AI部署工程师
      - AI算法工程师
    - 企业应用开发要求：企业级应用的高可用性
      - 分布式集群部署：负载均衡（nginx），避免单点故障
      - 容错和冗余设计：主备集群，自动切换
      - 监控和告警：便于即时恢复
    - AI应用开发和传统软件开发的区别
      - baseline迭代思维
      - 应对AI应用的不确定性
    - 项目管理： 企业里代码规范和代码管理
      - GIT
    - 自我提升：如何自我学习，跟进前沿技术
      - 如何快速成为一个新领域的专业人员
      - 如何跟踪前沿技术
    - AI岗位面试技巧

Presented with xmind