01大模型业务流程进阶20250419

- 大模型业务介绍
  - 模型的级别

特点	通用模型 L0	行业模型 L1	场景模型 L2
操作方式	直接使用开源、闭源模型	通用模型（基础模型）使用数据集进行微调后的效果	针对某一个特殊场景进行设计的模型

  - 开源模型与闭源模型选型特点
    - 闭源模型：
      - 效果好，维护成本相对较低，需要购买模型的服务，成本较高
    - 开源模型：
      - 模型脚本和权重可以免费获取，需要自行完成模型微调和部署的工作，对于技术能力要求高。
  - RAG（信息检索增强）
    - 通过给大模型添加知识库，可以帮助我们提升大模型对专业知识回答准确度一个效果
- 大模型业务流程
  - 业务场景需求分析
  - 模型选型
  - 微调数据准备
    - 如何获取高质量的数据
      - 数据来源根据实际业务场景进行获取
      - 对数据进行相关的清洗和处理，保证数据质量
      - 制作数据集时，要保证数据集规模（10k-100k）
      - 相似性：尽可以和实际业务中对话相似
  - 训练调优
    - 全参微调
      - 不建议使用，需要大量的计算资源进行处理
    - 高效参数微调（PEFT）
      - 冻结模型参数中大部分的参数，只调整小部分参数，效果与全参微调近似
      - Prefix Tuning
        - 前缀微调，会针对Embedding，Transformer前添加额外前缀参数，调整这部分参数进行处理
      - Prompt Tuning
        - 前缀微调简化版本，仅针对Embedding进行处理

        - P-Tuning
            - V1，V2
            - 加入MLP，LSTM网络层，进行更好的处理判断
        - LORA（最优）
            - 对原始模型参数进行低秩分解方式，获取关键参数进行微调
            - 可以泛化到全参微调的场景
            - 可插拔，根据实际场景进行切换
            - 原理简单
            - 无额外参数增加，不会增加推理时延
- 推理部署
    - 量化
        - 降低运行时计算设备的使用量，提升推理速度（保证模型能够项目的最低评估指标）
        - 量化、模型蒸馏、模型剪枝
    - 部署
        - 云部署
            - 大规模参数的大模型，提供大量计算资源。受限于带宽，会影响推理速度
        - 端部署
            - 适合移动端，信号接受难得场景中，受限于端侧设备的性能
        - 云端协同
            - 先试用端侧进行预处理操作，降低云测的运行压力，再将预处理后的数据交给云侧进行计算
- 应用集成