田海林

求职意向：大模型应用工程师 | 杭州，深圳
15185849917 | HailinTi@163.com

专业技能

- 熟悉 **模型蒸馏、长文本技术、RAG检索增强技术**，熟悉 **SFT** 、**RLHF** 等调优技术；
- 熟悉 **Embedding** 原理，熟悉 **openai** 等大模型 **api** 接口的调用；
- 熟练 **词向量模型**构建，如 word2vec，Char2vec，FastText，cw2vec等；
- 熟悉 **文本相似度**计算、**情感分析**、**文本分类** 等技术；
- 熟悉 **关系抽取、知识存储、知识图谱推理、语义搜索** 等知识图谱技术；
- 熟悉 RNN、LSTM、Attention、Transformer、Bert等深度学习框架；
- 熟悉 **Hugging Face** 生态，熟练调用各类训练模型与工具库；
- 熟悉 **Docker** 容器化技术，可实现模型的快速部署，掌握 **Kubernetes** 进行集群管理；
- 了解 AWS、阿里云等云平台，能在云端完成模型训练与部署任务；
- 能够调节 GPT 实现**智能画图**、**智能客服**等应用，设计过基于 OpenAI 接口的文档问答机器人；
- 部署过ChatGLM-6B、Llama3 等大模型，并进行训练、微调、量化；
- 对分布式 **DeepSpeed** 大模型分布式训练框架有所了解；
- 了解 LangChain，LangGraph， MCP，TransFormers agent， AutoGPT， ModelScope-Agent 等工具框架；
- 了解 图像分类、图像目标检测算法，如RCNN、SPPNet、Fast-RCNN等；
- 熟悉 MySQL，对 SQL 优化、索引、日志、主从同步、分库分表等有较为深入的研究。

项目经验

基于RAG的法律条文智能助手

- **项目背景**：
  构建法律领域智能问答系统，解决传统 LLM 对 “用人单位 / 劳动者解除合同” 等主体识别模糊、条款检索准确率低的问题，实现劳动合同法、工伤认定等场景的精准法律咨询。
- **核心技术工作：**
  1. **RAG 基础架构设计与实现：**
     搭建**检索增强**生成框架，集成**HuggingFaceEmbedding（BGE-small-zh-v1.5）**向量模型与**Qwen2.5-7B-Chat**大模型，实现法律条款的语义检索与回答生成；
     设计 “用户提问 - 向量检索 - LLM 生成” 三阶段流程，完成基础查询引擎开发，初始实现 Top3 条款检索功能。
  2. **检索策略深度优化：**
     提出 “初筛 Top10 + 精排 Top3” **两阶段检索机制**：初筛阶段通过向量相似度快速获取候选集，精排阶段引入 BGE-reranker-large-zh-v1.5**交叉编码器**模型，计算 query 与文档的语义相关性分数；
     优化评分机制：对比**双编码器**与交叉编码器的差异，将 “用人单位解除合同” 等法律主体的识别准确率提升**37.5%**。
  3. **高性能推理服务构建：**
     集成**vLLM**推理引擎，通过**--tensor-parallel-size 2** 配置实现 GPU 并行计算，将生成速度从 45 tokens/s 提升至 320 tokens/s（**7.1 倍**），显存占用降低 **38%**（7B 模型从 8.2GB 降至 5.1GB）；
     开发**vLLM-OpenAI API**接口适配层，替换原有 LLM ，将长文本响应时间稳定在 4.2 秒左右，满足实时问答需求。
  4. **可视化系统与工程落地：**
     使用**Streamlit**开发交互界面，实现法律问题输入、结构化回答展示及条款溯源功能（支持条款标题、来源文件、内容摘要的折叠查看）；
     完成系统全流程集成：部署**向量数据库**、RAG 引擎与 vLLM 服务，通过nohup命令实现后台稳定运行，典型案例（如 “试用期辞退补偿”）查询命中率达 100%。
- **项目成果：**通过优化调整，最终条款召回率达到**87.3%**，推理性能提升至**320 tokens/s**，显存占用（7B模型）降低至**5.1**GB

基于RAG的公司年报解析智能助手

- **项目背景：**
  构建企业年报智能问答系统，解决传统LLM在处理“公司年度财务报表解析”、“跨公司业绩对比”等复杂文档分析时面临的主体识别模糊、数据检索准确率低的问题，实现基于企业年度报告的精准信息提取与高效问题回答，涵盖财务数据分析、产品发布详情及领导层职位变动等多场景的专业咨询服务。通过集成先进的PDF解析技术、定制化的数据库管理方案以及语义搜索和大型语言模型重排序机制，本项目显著提升了对企业年报中关键信息的识别精度与响应速度，为用户提供了一个强大的工具来快速获取所需的企业资讯。
- **核心技术工作：**
  1. **PDF解析与处理：**选择 **Docling** 作为PDF解析工具，针对特定需求进行了源代码级别的修改，使得其能够生成包含所有必要**元数据**的JSON文件，并进一步转换为**Markdown文档**，以保留表格结构并改善格式。
  2. **数据库创建与检索优化：**
     针对每家公司创建独立的**向量数据库**，并利用 **FAISS** 进行数据存储和搜索。为了提升检索质量，引入基于**LLM**的**重排序**技术，结合初始的向量搜索结果，进一步筛选出最相关的页面。
  3. **增强与生成：**
     设计并实现了一个**模块化**的提示存储方案，以及针对不同类型问题的**路由机制**。应用**链式思考**(CoT)方法提升答案质量，并使用结构化输出(SO)规范模型响应格式，以确保准确性。
- **项目成果：**
  1. 成功构建一个高效的问答系统，在解析速度、检索精度和答案格式标准化方面展现出显著优势。
  2. 解决了PDF解析中的复杂问题，优化了从大量文档中提取关键信息的过程，提高了系统的透明度和可调试性。
  3. 实现了对复合查询的支持，能够准确地将复杂问题分解为简单子问题，并通过标准流程逐一解答，最终整合成完整答案。大幅减少了模型幻觉，提升了整体回答质量。

工作经历

海南汇英信息技术有限公司

- 时间：2024 年 6 月——2025 年 6 月
- 部门：技术部
- 职位：大模型应用开发工程师
- 工作内容：参与模型的设计与训练，协助团队将模型集成到实际应用中，并进行必要的调整和优化，确保在特定应用场景下的有效性和可靠性

教育背景

海南师范大学 - **软件工程** - **本科** 2020-09 ~ 2024-07

获得荣誉：在校期间，多次参加互联网+、蓝桥杯和程序设计竞赛并获得校级奖项。
主修课程：C语言程序与设计、Python语言、Pytorch框架、深度学习基础、计算机网络、数据结构、数据库原理等

荣誉证书

- 人工智能高级算法工程师证书
- 企业系统工程师(NIIT)
- 英语四级证书

自我评价

- 人工智能专业背景，具备扎实的深度学习理论基础，熟练掌握**PyTorch，TensorFlow**等主流框架
- 对新技术充满热情，时刻保持对**AIGC**前沿技术的敏锐度，持续学习多模态融合等知识，不断提升模型应用的创新型与实用性
- 擅长基于LLM进行模型微调与推理优化
- 注重**团队合作**，能将业务需求高效转化为技术方案