AI大模型应用开发工程师

姓名：刘均益 | 性别：男 | 年龄：25
学历：本科 | 工作经验：4年 | 专业：计算机科学与技术

教育背景

2018.9 - 2022.06 郑州工商学院 计算机科学与技术

工作经历

2024.03 - 2025.06 **杭州博彦信息有限公司** **大模型应用开发工程师**

- **背景：** 探索预训练语言模型在特定领域的应用及优化方法
- **职责：** 研究大模型针对垂直领域的微调技术，开发高效参数高效微调方法
- **难点：** 在有限的计算资源下提升模型性能，解决中文大模型的知识幻觉问题
- **成果：** 主导开发了多个ai项目，解决90%的模型选型难题，基于项目需求精准匹配开源模型。为项目前期制定数据规则，极大提高语料质量。主导模型微调，建立RAG本地知识库。合理制定量化方案平衡模型的精度与性能，节约30%的算力资源。

2021.12 - 2024.01 **杭州建海科技有限公司** **数据采集工程师**

- **背景：** 为公司项目后台数据库数据采集大量可靠的资讯类数据
- **职责：** 负责上百个站点的数据采集脚本开发，维护和数据清洗入库
- **难点：** 如何提高脚本的稳定性，确保采集工作的正常进行
- **成果：** 个人采集公司数据库90%的数据，为项目的正常运行提供有力支持

项目经验

2024.04 - 2024.08 **基于Qwen+Lora的ai试题系统** **llamafactory+qwen+vllm+lora**

- **背景：** 开发一款ai试题系统，只需要输入题目，ai自动提供答案和解题思路
- **职责：** 负责模型选型、训练策略制定、制定数据方案、评测和优化
- **技术难点：** 平衡模型规模与性能，解决模型理解专业领域知识的能力，优化推理速度
- **成果：** llamafactory+lora微调qwen模型，显著提高模型理解能力；vllm部署融合后的模型，相比同等规模模型推理速度提升60%；人工评估模型回答准确率97%

2024.09 - 2025.01 **基于RAG的法律条文助手** **llamaindex+qwen+chormadb+lora+rerank**

- **背景：** 企业客户需要能够理解法律领域知识，结合RAG检索本地知识库回答专业问题的AI助手
- **职责：** 设计基于大模型的知识增强系统，解决知识检索、融合与推理问题
- **技术难点：** 处理专业领域长文档理解，提升模型在专业知识上的推理能力
- **成果：** 开发了支持100万级知识条目的检索增强生成系统；设计了两阶段检索-生成架构，将专业问题准确率从65%提升至92%；

2025.02 - 2025.06 **基于RAG的智能客服系统** **llamaindex+qwen+chormadb+embedding模型**

- **背景：** 基于客户需求开发结合公司数据的线上智能客服。
- **职责：** 研究大模型压缩技术，优化推理性能，实现端侧部署
- **技术难点：** 在保证模型性能的前提下实现最大化压缩，平衡速度与质量
- **成果：** 提出混合精度量化方法，模型大小减少75%；开发注意力机制优化算法，推理速度提升3倍；

技术专长

- 精通Python语言，Transformer，huggingface，pytorch，Langchain，Langgraph等。
- 熟悉开源NLP模型，GPT，BERT，DeepSeek，Qwen等。
- 熟悉开源ai智能体搭建平台Dify，Coze，AutoGen等。
- 精通模型微调，Lora，QLora量化技术，精通微调框架Llamafactory，Xturne等。
- 精通模型部署推理优化，精通vllm，LMdeploy，huggingface，ollma等。
- 精通RAG技术，掌握RAG框架Llamaindex，RAGFlow。
- 擅长模型量化、知识蒸馏、参数高效微调等技术，能将大模型高效部署到各类场景。