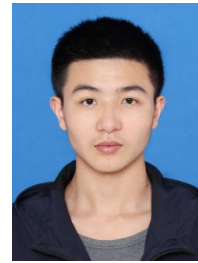


杨峰



求职意向：大模型应用开发工程师 / AI 后端工程师

☎：15575501892 ✉：code of yang@163.com ⚙：中共党员

教育背景

2023.9–2026.6 上海理工大学 硕士 系统工程/人工智能理论与应用
2019.9–2023.6 武汉纺织大学 学士 软件工程

实践实习

- 2025.07-10 上海航翼网络科技有限公司 产品研发 AI 研发实习生
- 异构环境推理优化：**负责 Qwen32B 模型在 x86 及 ARM 架构服务器上的环境搭建与适配。独立完成模型层级剪枝与 **AWQ 量化**，利用 **vLLM** 优化显存管理，在高并发场景下显著提升吞吐量并降低推理延迟。
 - 异步推理架构设计 (FastAPI + RabbitMQ)：**设计并实现基于 **RabbitMQ** 的异步任务处理系统。通过消息队列解耦前端请求与后端 GPU 推理服务，配合 Redis 实现任务状态追踪与削峰填谷。
 - 部署交付：**编写 Dockerfile 对推理服务进行标准化封装，并部署至 L20 算力集群进行压力测试。
- 2025.03-2025.06 上海同梁智能科技 算法研发 AI 算法实习生
- 数据流水线搭建：**基于 Python 编写自动化爬虫与 ETL 脚本，利用 Tushare 接口获取金融数据，使用 Pandas 进行清洗（缺失值填充、因子处理）并存入 MySQL 数据库。
 - 模型优化：**参与**多智能体对抗学习模型 (MAA)** 的迭代，设计特定的损失函数与分类指标，通过调整时间窗口与反归一化层参数，优化了模型的预测性能。

项目经历

- 2025.07–2025.09 基于大模型的智能代码助手插件 负责人
- 项目描述：本项目是一款面向 ECU 开发者的 VSCode AI 插件，基于 Qwen 模型微调，深度适配 MISRA C/C++ 安全编码标准。
- 垂直领域数据工程：**针对工业级安全规范 (MISRA C++ 2012)，设计半自动化数据清洗脚本，人工校验并整理 **4200+** 条高质量指令微调数据，构建了包含“违规代码-合规修正-规则解释”的结构化数据集。
 - 模型微调与评估：**基于 **Qwen-32B** 基座模型，采用 **LoRA** 技术进行领域适配训练。通过调整 Rank 和 Alpha 参数平衡训练效率与效果，最终在 MISRA 规则识别与修复任务上，将准确率提升至 **92%+**。
 - RAG 系统开发：**针对长文本规范文档，构建了基于 LangChain 的检索增强生成 (RAG) 模块，实现了代码上下文感知 (Context-aware) 的交互式问答，有效缓解了模型幻觉问题。
 - 工程落地：**开发 MISRA 合规检测模块，集成微调后的推理接口，实现代码自动审查与修复，修复效率较人工提升 **60%**，显著降低了 ECU 软件的功能安全认证成本。

个人技能

- 大模型工程化：**深入理解 Transformer 与 vLLM 原理，熟练掌握 **SFT (LoRA/QLoRA)** 微调及 **RLHF** 对齐流程；熟悉模型剪枝与 **AWQ/GPTQ** 量化技术。
- RAG 与 Agent：**熟练使用 **LangChain** 构建 RAG 应用，掌握向量数据库 (Milvus/Faiss) 及检索增强策略；熟悉 **CoT (Chain of Thought)** 及 Few-shot 提示词工程。
- Python 后端：**精通 Python 异步编程，熟练使用 **FastAPI** 开发高并发推理接口；掌握 **RabbitMQ + Celery** 构建异步任务队列；熟悉 MySQL 索引优化与 Redis 缓存策略。
- DevOps：**熟练使用 Docker 容器化部署；熟悉 Linux 环境及 Git 工作流；了解 K8s 基础运维。

所获奖励

- 2021-2022 获国家励志奖学金 (2 次)、主持省级大创项目，参与国家级大创项目、“优秀学生干部”
- 2023-2026 **第二十二届“华为杯”二等奖**、第四届上海市高校开放大数据分析挑战赛一等奖、全国大学生创业决策仿真大赛一等奖、上海理工大学“挑战杯”大学生创业计划竞赛校赛二等奖