

李赢洲

电话: 13054111853 | 邮箱: 13054111853@163.com
GitHub: github.com/20082123 | 个人网站: my-flow.top
求职岗位: AI 应用开发工程师 | 2027 届



教育背景

贵州大学 (211 / 双一流) 2024.09–2027.06
电子信息 (硕士, 2027 届); 研究方向: 边缘计算、强化学习与随机配置网络; 软件设计师 (中级); 研究生数学建模国家二等奖
防灾科技学院 2020.09–2024.06
信息管理与信息系统 (本科); 河北省数据分析竞赛省三等奖; 校二等奖学金 2 次

实习经历

上海波士内智能科技有限公司 AI 应用工程实习 | 2026.01–2026.04
参与工业文档智能问答 PoC 与电梯故障诊断项目, 负责文档解析、异步任务、RAG 模型调用优化、图谱入库和离线评测。
• **文档数据链路**: 针对文本页、扫描页及复杂表格混排问题, 设计“逐页检测 + 动态路由”解析链路, 使用 pdfplumber 与 PaddleOCR 按页处理, 将 7 页混合 PDF 解析由分钟级缩短至十余秒; 使用 UNWIND 完成 Neo4j 节点与关系批量写入。
• **异步服务**: 将长文档处理改造为 FastAPI + SQLite + Worker 异步任务, 提供上传、状态查询、失败记录与重试接口, 避免 OCR 及模型长任务持续阻塞 Web 请求。
• **模型调用优化**: 针对 LightRAG 请求过程中重复初始化本地模型的问题, 将模型客户端改为进程级常驻复用, 使冷启动仅发生一次、后续单 Chunk 调用稳定在秒级, 并避免多实例加载导致的显存增长与 OOM 风险。
• **模型评测**: 审查工单、运行日志与维修记录的数据可见时间, 识别目标泄漏并建立 68.4% 的可靠基准; 编写 Python 脚本完成 1,500+ 样本并发评测、JSON 清洗与 Bad Case 归因。

项目经历

面向智能座舱离线场景的端侧 LLM 助手
个人项目; 技术栈: Python、FastAPI、llama.cpp、Qwen2.5-1.5B、LoRA、sherpa-onnx、ONNX Runtime、SQLite FTS5/sqlite-vec、systemd。
• **端侧部署**: 在树莓派 5 (4GB) 全离线环境中部署 Qwen2.5-1.5B Q4 模型与 bge-small-zh INT8 向量模型, 使用 FastAPI 构建统一入口; Q4 生成速度达到 11.77 tok/s, 相较 Q8 提升 79%, 模型体积缩小 40%。
• **评测驱动**: 将 464 页车主手册处理为 1,387 个切片, 基于 jieba+FTS5 与 sqlite-vec 构建 BM25/向量双路检索并使用 RRF 融合; 40 道有答案题中 Hit@5 达到 82.5% (BM25 70%、纯向量 80%), 批量评测发现并回滚 KV 缓存复用导致的答案污染。
• **微调闭环**: 构造 2,690 条训练样本, 并通过独立模板池生成 200 条同源测试样本 (150 条车控、50 条范围外); 完成 LoRA 微调、GGUF Q4 量化与树莓派部署, 车控意图与槽位全对率由 16.7% 提升至 94.7%, 范围外误触发率为 6%。
• **语音交互**: 基于 sherpa-onnx 实现本地唤醒与语音识别, 结合能量 VAD 和本地 TTS, 在 USB 麦克风、扬声器上实机打通“唤醒—识别—意图路由—回答—播报”全离线链路; 根据现场口语测试中暴露的领域词误识别、问句误路由和长回答播报等待问题, 引入热词偏置、问句保护与逗号分句播报。

面向车联网边缘计算的轻量化在线任务卸载 2024.10–至今
核心研发; 论文第一作者; 技术栈: Python、PyTorch、强化学习、随机特征网络、树莓派 5。
• **算法设计**: 面向无人机辅助车联网中的在线卸载问题, 设计采用冻结随机映射层与可训练 Dueling 输出头的 RF-DQN 轻量化值网络, 将可训练参数由 34,052 降至 1,024; 推理与单次 TD 更新的线性计算量分别降低约 18.8 倍和 26.2 倍。
• **实机验证**: 在树莓派 5 上完成部署测试, 推理延迟由 0.51 ms 降至 0.14 ms, 单步更新延迟由 5.41 ms 降至 2.10 ms, 模型文件缩小约 13 倍, 并在三类移动场景中保持有效卸载性能。

论文成果

SCI 一区 第一作者论文《面向资源受限车载端的轻量化在线任务卸载方法》投稿 *IEEE Internet of Things Journal*, 返修中。
北大核心 第一作者论文《基于 Huber 损失和流形正则化的 SCN 鲁棒半监督分类方法》被《计算机工程与应用》录用, 预计 2026.12 见刊。

专业技能

Python 后端 熟悉 Python、FastAPI、asyncio、RESTful API、SQL、MySQL/SQLite、Git, 具备异步任务、接口开发、数据处理和性能测试实践。
AI 应用 熟悉 RAG/LightRAG、Embedding、FAISS、SQLite FTS5/sqlite-vec、Neo4j; 具备 LoRA 微调、llama.cpp/ONNX Runtime 端侧部署、离线评测与 Bad Case 分析实践。
算法与实验 熟悉 PyTorch、强化学习与随机配置网络, 具备仿真环境搭建、对比实验、性能测试及树莓派边缘部署实践。
基础能力 掌握 Linux、Docker、systemd 基础操作, 了解 C/C++; 熟悉常用数据结构、计算机网络和操作系统基础; 大学英语四级。