

教育背景

墨尔本大学 软件工程 硕士
四川大学 网络空间安全 本科

2024.07 – 至今
2020.09 – 2024.06

专业技能

- 语言: 熟练掌握 Java、Python 等编程语言、了解 JavaScript/TypeScript, C/C++ 等语言
- 后端: 熟练掌握 Spring Boot、FastAPI 等主流后端框架; 掌握 IoC、AOP 原理及微服务架构; 熟悉 Java 并发 (synchronized、volatile、ThreadLocal、线程池、锁机制) 和 Python 异步编程 (async/await、协程)。
- 数据库与中间件: 熟悉 MySQL, 了解索引原理, 优化方式及事务控制和锁机制, 了解中间件 Redis、Kafka 等
- AI/LLM: Transformer、Pytorch、LangGraph、RAG、Multi-agent、prompt-engineering、openclaw、Claude Code (skills, coworker)

工作经历

北京声智科技有限公司 全栈开发工程师 (实习) 2025.06 – 2025.07

- 负责 iOS 客户端 + Java 后端的核心功能开发, 打通耳机语音采集 → 实时翻译 → App 展示的完整链路。
- 设计并实现会员等级与权限体系, 控制实时翻译、语音转写、历史记录等能力按会员等级动态开放。
- 后端基于 Spring Boot + MySQL 实现用户身份、会员状态与 API 鉴权, 支撑高频语音请求。
- 使用 WebSocket 构建实时语音服务, 实现语音流的低延迟传输与翻译结果实时回传。

北京麒纪科技有限公司 AI 应用开发工程师 (实习) 2026.01 – 2026.03

- 使用 Python + FastAPI 搭建乒乓球赛事 AI 解说系统, 实现视频输入 → OCR 识别 → LLM 生成 → TTS 合成 → 视频融合的完整链路, 支持 RTMP 直播流实时处理。
- 构建球员、赛事结构化数据模型, 针对不同轮次差异化设计; 开发 TTS 读音修正模块与实时记忆追踪, 确保解说连贯无重复。
- 设计预赛备稿 Agent 融合球员档案、历史交手等数据生成叙事框架, 采用增量上下文融合机制将备稿注入解说生成; 预留赛中动作识别 Agent 接口, 与运动员识别模块解耦合支持直播实时决策。

项目经历

NanoMemAgent - 带记忆模块的轻量化 AI Agent 执行框架 2026.02 – 2026.03

Python + ReAct + Skill + Tool + MCP + ChromaDB + Neo4j + BGE-M3 + BM25 + ThreadPool

项目描述: 从零实现一个模块化 Agent 执行框架, 内置 ReAct 推理循环、动态 tool 调用和 MCP 接入、设计分层记忆系统 (短期向量存储 + 长期知识图谱) 及插件化 Skill 机制, 支持多模型切换, 并且已接入飞书。

- Agent Loop 核心引擎:** 实现 NanoMemAgent 心脏模块, 协调消息路由 → ContextBuilder 上下文组装 (系统提示 + 记忆 + 技能) → LLM 推理 → Tool Calls 执行 → 结果反馈的完整循环, 支持最大迭代次数保护与工具链式调用。
- 分层记忆系统:** 设计 L1-L5 五层架构 (存储/嵌入/检索/知识/应用), 支持 Operational Memory (动态 CRUD) 与 RAG Knowledge Base 双轨存储; Vector + BM25 + Graph 三路并行检索 + RRF 融合排序。
- 模块化工具生态:** 实现 MCP Client 标准化封装, 支持文件系统、数据库、搜索等外部工具动态接入; 内置 Cron 定时任务调度、子代理 (Subagent) 后台执行、Heartbeat 健康检测等核心能力, 抽象 LLMProvider 接口层统一封装 OpenAI、Anthropic、本地模型等多渠道零代码切换。

PocketLLM 极小语言模型全链路复现 2025.09 – 2025.11

PyTorch + Transformers + MoE + SFT + LoRA + DPO + GRPO + Flash Attention

项目描述: 针对单卡算力受限环境下的模型轻量化需求, 从零构建一个 0.2B 参数规模的极小模型, 完整实现预训练、监督微调 (SFT) 及多种偏好对齐 (RLHF/RLAIF) 算法在内的完整链路。

- 全链路训练:** 独立开发结合 MoE 的 Decoder-only 核心架构 (512 维度)。完整实现了预训练、基于 LORA 的高效指令微调 (SFT), 以及涵盖 DPO/PPO/GRPO 的偏好对齐训练全流程。
- 数据工程与处理:** 使用清洗后的 MNBVC 语料进行预训练; 针对 SFT 阶段, 利用 COIG 构建高质量短文本 (<512 字符) 数据集; 设计规则消除噪声, 显著提升模型指令遵循能力。
- 基准评测:** 在 C-Eval、CMMLU 等中文基准测试中, 该 0.2B 模型在有限算力下取得了与同量级基线模型相当的准确率 25% 左右的准确率。