周馨蕊

(+86)183-9537-8121 · xinruizhou2002@gmail.com · https://xinrui-z.github.io/ ·

教育背景

安徽理工大学, 数字媒体技术, 工学学士 2020.09 - 2024.06

- 主修课程：人工智能、计算机视觉、计算机图形学、数字图像处理等。

工作经历

上海胜算速惠云科技有限公司 | 大模型部署工程师 & 产品技术负责人 2025.07 - 至今

- 负责 Qwen-Image、Flux、LoRA 等模型的端到端部署与推理优化，基于 PyTorch、Cog、ComfyUI 设计高性能推理架构，显著降低延迟并支持快速版本迭代。
- 构建大模型 API 全生命周期管理机制，基于稳定性、吞吐量、资源利用率等指标动态调度，上架 GPT、Claude、Qwen 等系列模型，支持日均调用 10K+。
- 建立标准化模型评测与选型体系，撰写技术评测报告，为内部决策与客户交付提供依据；推动开发者生态建设，提供微调、部署、API 接入与 Prompt 优化技术支持，服务超 1000 名开发者。

杭州澜浔科技有限公司 | 联合创始人 & COO 2024.07 - 2025.01

- 独立完成小程序核心模块开发，主导产品从 0 到 1 研发及迭代，深度参与用户调研和需求分析，50 天内完成产品上线，一周注册用户 1000+。
- 搭建公司商务体系，基于调研与竞品分析明确用户群体和商业化路径，为后续增长奠定基础。
- 建立数据驱动决策体系，通过用户行为分析推动功能迭代，核心功能日活提升 30%，显著优化产品体验与活跃度。

之江实验室 | 算法工程师 - 实习 2023.12 - 2024.05

- 调研并复现 3D 生成、重建和渲染前沿算法，追踪 CVPR、ECCV 最新研究，分析 20 余篇论文，复现核心算法并提供技术实现方案。
- 参与 3D 生成与重建相关项目，基于 PyTorch、Stable Diffusion 实现模型构建及性能优化，并参与撰写 3 篇科研论文。

项目经历

浔之驿站小程序 | 研发 & 产品负责人 2024.08 —2025.01

- 主导产品从 0 到 1 的规划与核心功能设计，结合用户调研和市场分析绘制用户旅程图；输出产品原型与设计稿，优化端到端体验并形成完整需求文档。
- 完成 Qwen 大模型适配与 Prompt 模板设计，调试优化 AI 对话逻辑以匹配业务需求。
- 独立开发高复杂度星盘 API，设计多维度运算与并发架构以支撑复杂业务逻辑，确保在多场景下数据处理的高稳定性与低延迟；基于 Spring Boot + MyBatis Plus 构建后端，并结合 Redis 显著提升系统并发处理能力与响应速度。

基于 AIoT 的煤炭输送机智能巡检装置 | 项目负责人 2021.10 —2024.05

- 主导智能巡检装置从需求调研、架构设计到边缘部署全过程，实现矿井下煤矸石高精度检测，并确保模型低成本稳定运行。
- 设计基于深度残差网络的轻量化架构，结合模块化重构、剪枝与蒸馏，大幅压缩模型体积和计算量；完成前后端推理加速与性能优化，构建端到端可部署方案。
- 成果：设备在湖北某电厂稳定运行 2 年以上，申请发明专利 2 项、发表 SCI 论文 2 篇。

技术能力

- **大模型部署与优化：** PyTorch、vLLM、Cog、ComfyUI、TensorRT
- **轻量化与微调策略：** LoRA、PEFT、Adapter、剪枝、蒸馏、低资源部署、多任务微调
- **模型评测与 Prompt Engineering：**模板设计与调优、标准化评测、Agent 开发、业务场景定制
- **工程与产品能力：** C++、Python、Java、Vue、React、Spring Boot、Redis、Figma

个人荣誉

- 安徽省计算机学会 2024 年度优秀毕业论文（19 人评选）
- 第 15 届中国大学生计算机设计大赛银奖（AI 方向）