

郑其昌

☎ +86 18862132686 ✉ qichang.zheng2000@gmail.com 📍 上海
📞 18862132686 📧 github.com/QichangZheng
👤 在职



工作经历

奇绩智峰智能科技有限公司-AI 技术负责人

2026年03月 - 至今

- 公司最高营收项目的技术负责人（团队 10 人），交付 AI Agent 全栈平台并通过互联网与政务等多套生产环境验收，实现约 5000 万元年营收；**主干累计交付 256 项功能特性、沉淀 280+ 篇设计文档，个人合入 680+ 个 PR。**
- AI Agent 构建和运行平台（North Agent Cloud）**
 - 主导 AI Agent 云平台研发**：建成覆盖调度编排、沙箱隔离、出网管控、凭证托管、权限审计、外部接入、弹性伸缩、异构 Harness 兼容、模板和技能市场、自然语言构建 Agent、对象存储挂载、全链路追踪审计等核心能力的平台底座。
 - 手脑分离优化沙箱生命周期和调度**：针对 ReAct 范式下 LLM 调用长时低负载和工具执行短时高负载的特点，**将运行时与沙箱分离**，并通过负载均衡和弹性伸缩提高资源利用率和并发性能；**以懒启动和自动回收将单位资源承载的沙箱数提高 10x。**
 - 平台能力闭环**：轨迹分析、对话构建 Agent 等平台内置核心功能，本身就是部署在本平台上的 Agent。
- 多 Agent 并行开发平台（Agent Team，独立开发，1400+ commit / 28 万行）**
 - Agent 协同与长任务稳健**：按模块拆分 Agent、各自独立 worktree 并行开发并相互通讯，**以专一的上下文和端到端独立对抗验证来提高大型复杂项目的质量**；通过巡检与**自研 hook** 治理长程任务的典型失效（说了要做却没做、任务僵死等）。
 - 云端调度与可观测**：控制面统一调度多机器、多 Harness（基于 Claude Code，兼容 Codex）；内建状态看板、巡检和自愈、异常实时上报与远程干预、PR 跟踪与多集群监控、定时任务与事件订阅；提供网页、iOS、飞书、CLI 四个入口。上线后日常并行 5 至 10 个 Agent（峰值 20+），**个人月均合入 PR 由 32 增至 212、交付功能由 8 增至 30 项。**
- 高性能沙箱服务（North Sandbox）**
 - 基于 E2B 协议实现沙箱平台**：涵盖预热池、租户体系、模板配置、网络策略、凭证隔离等功能供多方使用。
 - 提升沙箱响应速度**：基于沙箱模板创建预热池，实现**沙箱毫秒级认领（提速 20x）**，大幅提高 Agent 启动和执行速度。
 - 分布式文件系统提高效率 and 扩展性**：使用 JuiceFS 连接对象存储作为沙箱文件系统，实现数据实时落盘和存储卷灵活切换。
- 网络防火墙和凭证隔离服务（North Firewall，一人从零建成）**
 - 内核级流量拦截和请求审计**：基于 iptables 透明拦截出网流量，并通过 CA 证书注入和 MITM 解析审计 HTTPS 请求。
 - 凭证隔离**：通过 Vault 保护凭证，并以占位符形式注入环境变量，Agent 只接触到占位符，凭证泄露面收敛到单一网关。

Devz AI-AI Scientist

2024年08月 - 2026年03月

- 基于 ReAct 范式的企业级 Agent 平台**：私有化部署并开发 30+ Agent 和 50+ 配套可复用的工具，实现覆盖意图识别、方案选择/分析/生成、任务规划/执行、结果验证的端到端自动化，支撑约 \$6M/年的订单。
- 解决 Agent 在复杂网页环境中的操作难题**：开发基于 Selenium 的 Web Agent，结合 LLM 意图识别和 HTML DOM 树解析，通过递归交互方式，**实现对非结构化网页的精准操作**；并引入视觉模型验证操作结果，实现 63% 的端到端故障自愈。
- 设计动作空间缓存系统**：构建“**脚本动作对齐+变量填充+多路径缓存**”机制，实现从 runbook 一键生成动作脚本和单一脚本解决多个事故的高泛化能力，并通过复用历史成功的执行路径，实现 Web Agent 20x 执行加速。
- 垂类模型微调**：主导垂类 Ops LLM 的后训练 SFT。搭建数据管道（内部数据 + 基于文档合成的高质量 QA），用 Unsloth 加速训练、rsLoRA 稳定高秩微调，使模型在内部运维问答评测集上的准确率相对 GPT-4/Opus 提升 55%。

技能

- Agent 工程**：Agent 开发交付全链路，Runtime/Sandbox/Gateway 解耦，Harness 工程，沙箱生命周期管理，E2B 协议兼容
- 编程与云原生**：Python(LeetCode 1500+)/Rust/TypeScript, Kubernetes, Helm 多环境 overlay/GitOps, KEDA 弹性伸缩, Docker/多架构镜像 (amd64+arm64), PostgreSQL/Redis Cluster, MITM 审计, Vault 凭证托管, 多租户 RBAC

教育经历

芝加哥大学-计算机社会科学 硕士（凤凰学术奖学金\$63,430，GPA：3.9）

2022年09月 - 2024年06月

西交利物浦大学-经济与金融 本科（杰出学术奖学金¥20,000，GPA：3.93，Rank：5/250）

2018年09月 - 2022年06月

Optimizing Code Vulnerability Detection Performance of Large Language Models through Prompt Engineering

- [文章链接](#) | 建立基于提示工程的代码漏洞检测优化框架，通过零样本/少样本及思维链（CoT）策略评测 GPT-4 等主流模型，针对 C++/Java/Python 代码库实现了 F1 分数最高 23.7% 的提升。

Domain Adaptation Analysis of Large Language Models in Academic Literature Abstract Generation

- [文章链接](#) | 主导跨学科 LLM 领域适应性研究，构建多维评估框架分析 2,400 篇学术摘要，量化了 GPT-4 等模型在计算机（0.847 分）与社会科学（0.623 分）领域的性能差异，揭示了关键语言特征对模型迁移的影响。

An Empirical Study on the Accuracy of Large Language Models in API Documentation Understanding

- [文章链接](#) | 开展跨编程语言的 LLM API 文档理解实证研究，建立包含自动化指标与人工校验的评估框架；评测 GPT-4、CodeT5 等 5 种主流模型在 Java/Python/C++/JS 上的表现，量化了 67.3% 至 89.7% 的准确率差异，为代码辅助工具的文档设计提供了关键优化路径。

Interactive AI agent for code refactoring assistance: A study on decision-making strategies and human-agent collaboration effectiveness

- [文章链接](#) | 提出一种基于时序图神经网络 (TGNN) 和强化学习的多层交互式 AI 智能体架构，通过 45 名专业开发者的对照实验，验证了该架构在代码重构任务中相较传统工具实现了 34.7% 的质量提升，并保持了 89.2% 的用户接受率。

An Empirical Study on the Quality Assessment of Code Comments Generated by Large Language Models for Different Programming Paradigms

- [文章链接](#) | 构建包含 450 个代码样本的跨范式评估框架，系统评测了 GPT-4 等 5 种 LLM 在面向对象、函数式及过程式编程中的注释生成质量。研究量化了范式间的显著性能差异（面向对象准确度 4.4/5.0 vs 函数式 3.5/5.0），揭示了 LLM 在处理高复杂度函数式代码时的能力短板。

HOW CAN WE USE CHATGPT BETTER: A RESEARCH OF API-ENHANCED CHATGPT IN STOCK PREDICTION 硕士毕设

- [文章链接](#) | 提出一种集成实时股价与新闻数据的 API 增强型 ChatGPT 架构，通过优化提示工程策略，使股价趋势预测准确率提升 10%，风险调整后收益（夏普比率）提升 20%。