徐 强

15291649017 | xuqiang5250@163.com | 3 年工作经验 | www.anyonghuan.me | GitHub: Qiang-Maker

AI 应用开发工程师 | AI 全栈工程师 | AI Agent 研发工程师 | FDE

个人优势

- **Agent 运行时与多智能体编排**：构建基于 **Agent Loop** 自主决策循环与「一切皆插件」**Harness 运行时底座**，落地 SubGraph 业务解耦与 **PostgreSQL Checkpointer** 状态机持久化及断点恢复；结合 **LangGraph / LangChain** 完成多智能体协作编排，串起「需求评估 → 计划 → 执行 → 复盘」的完整业务链路；落地多层长期记忆隔离与上下文治理，集成 **astream_events** 捕获 Token 级审计中间态，具备基于规则沉淀的自进化闭环机制。
- **垂直 RAG 检索与实时数据问答**：落地客运 SaaS 智能应答与工业知识库，打通「**RAG 文档检索 + 业务库实时查询**」混合问答链路；文档侧按章节语义切片并做**父子块（Parent-Child）双层检索**；构建「**BM25 专有名词召回 + HNSW 稠密向量（结合 Int8 量化）**」多路召回，通过**动态调节 dense_weight** 结合 **Cross-Encoder Rerank 精排与 MMR 去重**，配合 **pgvector** 索引调优，提升长尾问题召回准度并抑制事实性幻觉。
- **业务数据治理与大模型评测闭环**：参与客运业务数据治理与多轮会话清洗，落地**正则与 NER 双层脱敏**链路；参与垂直领域意图分类的 **SFT / QLoRA** 微调数据准备与 **DPO** 偏好对（chosen / rejected）构造；搭建基于 **LLM-as-a-Judge 与 Cohen's Kappa 一致性校准盲评**的双层评测集，结合 **Prefix Caching 与 KV Cache** 优化线上首字延迟（TTFT），建立线上 **Badcase 回流与检索评测闭环**。
- **结构化输出契约与模型调用治理**：在生产级问答与 Agent 链路中，运用 **with_structured_output、Pydantic Schema 与 JSON Mode** 落地强类型契约化输出约束；构建输入敏感意图拦截、工具调用白名单校验与业务规则二次对齐的**三层防御护栏**；在**应用层**搭建动态模型路由与 **Max Token 分级配额**计量，实施异常调用熔断与备选模型降级。
- **场站视觉中台与端云流式多模态交互**：参与**水陆场站安检视觉中台**建设，落地 **Qwen2-VL** 完成视频关键帧时序推理与工业/票据 **OCR** 结构化抽取；搭建端云协同的**流式语音链路**，整合端侧 **VAD 门控**与云端 **ASR / TTS** 级联，引入 **FEC 前向纠错与自适应 Jitter Buffer** 机制保障弱网语音可用性，打通「设备感知、视觉检测、语音播报、工单流转」的多模态作业闭环。
- **工业互联网与物联网设备对接**：参与设备对接设计，涉及 **Modbus（RTU / TCP）、MQTT、串口通信、TCP / UDP** 等协议，完成**设备数据采集与状态采集**；对接**安检闸机、人脸识别终端、工业机台 PLC、边缘采集传感器**等现场设备，落地点位映射、心跳保活、断线重连与底层报文解析，打通物理设备数据与业务后台、AI 中台之间的链路。
- **跨语言 AI 全栈交付与异步流式架构**：基于 **C#（.NET Core / Web API）与 Python（FastAPI / asyncio）**构建双后端服务，配合 **Vue3 + TypeScript** 与 Electron + React 完成 MES 前端平滑重构、线上工具站及多平台桌面客户端的全链路交付；在 AI 实时交互场景落地 **WebSocket 全双工**与 **SSE 流式推送**的容错选型，并设计**异步并发调度**、限流与 TraceID 全链路日志基建，保障生产可用性。
- **规约驱动的 AI 协同工程与全栈交付**：结合 **Claude Code、Codex、Cursor** 构建标准化 AI 协同开发流程，推行「需求规约化 → 开发计划拆解 → AI 按计划执行 + 人工审查」的交付模式；把架构约定与代码规范沉淀为可复用的 **Rules / Skill** 资产，独立完成多款线上产品从开发到上线维护。

个人作品 & 开源实践

01 Super Agent 智能体运行时 · Harness 自进化底座　TypeScript · Node.js · Harness · MCP　github.com/Qiang-Maker

自研企业级智能体运行时：构建基于 Agent Loop 决策循环与「一切皆插件」Harness 底座，支持 Tool 白名单动态注册与长会话多层记忆隔离；首创 Badcase 与人工修正信号自动捕获机制，自主沉淀为全新规则护栏与 Few-Shot，实现规则自进化生长。

02 Agent OS 多 Agent 协作工作台　TypeScript · 飞书开放平台 · MCP · Claude Code / Codex　github.com/Qiang-Maker

面向「一个人带一个 Agent 团队」的研发协作场景，以**飞书**为统一入口实现**多 Agent 协作工作台**：群内 @ 机器人发起任务，按职责角色拆分并分发，串起「**需求评估 → 计划 → 执行 → 复盘**」闭环；底层实现会话与任务持久化、跨轮上下文管理与复核轮次控制，并通过统一适配层接入 **Claude Code / Codex** 等多种执行引擎；飞书自建机器人已发布，权限与长连接配置完成，服务常驻运行中；**该工作台由我在个人探索中发起并沉淀，现已在团队内推广用于日常研发协作。**

03 群忆 · 微信共享相册小程序　微信小程序 · 线上运行产品　qunapi.goodssoft.com

针对群组活动后照片收集繁琐、画质被压缩且难以长期整理的问题，独立研发并上线运行微信共享相册小程序：支持**免安装即用**、**原图上传与保存**、点赞评论、**回忆页 Vlog 一键成片**，并按时间轴与照片集长期沉淀。

工作经历

倍施特科技（集团）股份有限公司

AI 应用研发工程师　2025.01 ~ 至今

- **智慧客运大脑建设**：主导「途小智」客运客服机器人大脑的应用侧建设，负责端云协同流式语音管道、垂直客规 RAG 检索链与多轮意图识别的落地交付；配合算法侧推进特定场景语料治理与双层自动化评测。
- **客运数据中台与客流态势预测**：沉淀客服多轮会话与站端预警日志，搭建清洗脱敏与标准化数据工程流水线；基于历史同期出行规律与现场过闸热度构建客流预测模型，向景区与枢纽输出节假日峰值及运力票务预测服务。
- **边缘设备接入与安检视觉分析**：负责站端刷脸机、闸机与客流感知设备的多协议接入；主导基于 Qwen2-VL 的安检视频流异步分析中台与任务失败模型设计，支撑水陆场站通行态势预警。
- **统一模型网关治理**：参与集团大模型统一调用网关设计，落地虚拟密钥体系、调用成本拆分计量与路由分级策略。

成都任我行软件股份有限公司 AI 全栈开发工程师 2024.03 ~ 2024.12

- **管家婆小工单 MES 全栈迭代：**负责企业级 MES 生产工单管理与开票打通；基于 C# / .NET 与 Vue 推进遗留架构向 .NET Core Web API + Vue3 的平滑重构，并推动构建发布与交付流程优化，提升服务迭代效率。
- **车间设备侧数据链路打通：**深入生产车间一线，对接加工机台与下位控制设备侧的数据通道，完成运行点位映射与状态回传，实现产线数据与 MES 的实时闭环。
- **「小辉智能体」车间排障多 Agent 系统：**从 0 到 1 打造面向车间运维的智能体系统，拆分对话、排障、知识库三大 Agent，引入 Checkpointer 实现长链路排障状态可追溯。

核心项目经历

一、「途小智」智慧客运语音机器人与 Harness 自进化交互系统

倍施特科技 AI 应用研发工程师 2025.01 ~ 至今

> 面对全国重点客运枢纽客服标准不一、退改政策更新频繁及高峰期排队严重痛点，主导「途小智」端云协同客服机器人大脑建设。业务覆盖移动端「团子出行」App 与站端公共服务终端，重点攻关老年人群体适老化交互与候车厅强噪声语音识别。本人负责 AI 应用侧全链路交付，直线上报技术负责人。集团客运 SaaS 平台公开转让说明书披露覆盖车站超 500 个、累计售票超 4000 万张，AI 客服作为平台核心模块随线推广。

· 阶段一：基础 RAG 知识库与 RTC 适老化语音链路优化

- **端云流式交互与适老化体验：**针对候车大厅强噪声与老年旅客语速慢、停顿长特点，优化端侧 VAD 门控（静音判定窗口动态放宽至 1.2s~1.5s）与云端高精度 ASR 级联；基于 WebSocket 双工流上传音频、SSE 增量下发文本，配合 TTS 声音复刻与语速调优（0.85倍+关键车次停顿强化）；协同优化 RTC 传输策略，配置自适应抗抖动缓冲（Jitter Buffer），在弱网环境下实现会话无感恢复。
- **专有名词加速与垂直 RAG 检索：**按语义逻辑切分客规与禁限带名录；构建「BM25 专有名词 + HNSW_HYBRID Embedding 稠密向量（结合 Int8 量化）」双路召回，引入 base-rerank / Cross-Encoder Rerank 精排并动态调节 dense_weight 平衡专有名词，存储降低 75%，**检索命中率由 72% 提升至 89%**；合并 System Prompt 消除哈希抖动，深度触发服务端 **Prefix Caching 与 KV Cache 显存复用，将复杂 RAG 首字延迟（TTFT）从 15s 压缩至 2s 内。**

· 阶段二：多分支复杂工作流编排、数据中台消费与 SFT/DPO 微调下沉

- **多分支复杂工单工作流编排：**面对票务退改、失物招领、特规行李审核等复杂场景，将简单单轮链路升级为基于 Plan-Execute-Replan 与 ReAct 的多分支 Agent 工作流；引入 **PostgreSQL Checkpointer 状态机**实现全状态持久化与任意中断节点断点恢复，并联动图文、视频多模态卡片受控混排输出。
- **消费中台数据与微调工程（微调管风格，RAG管事实）：**消费数据中台（项目三）清洗脱敏后的高质量对话；本地基于 Qwen2.5 完成 QLoRA 超参验证，上传云端完成生产级 LoRA 微调并引入 DPO 偏好对齐；建立基于 **LLM-as-a-Judge 与 Cohen's Kappa 一致性校验盲评**的双层评测机制，**人工盲评风格一致性评分从 2.1 跃升至 4.3（5分制）**；建立 Max Token 分级控制模型优化推理成本。

· 阶段三：Harness 运行时底座构建与 Agent Loop 自进化闭环

- **模型无关的 Harness 运行时底座：**从零构建「一切皆插件」的 Agent Harness 底座，抽象统一 LLMProvider 接口与工具注册表；基于 **MCP 协议**将算力路由、垂直检索与会话记忆解耦为独立共享服务，供全国场站一体机与移动端多终端无缝复用。
- **示范学习与话术规则自进化闭环：**依托 Agent Loop 决策循环构建长会话多层上下文治理；建立 Badcase 纠错反馈监听机制，自动捕获一线坐席“示范解答+点评式教学”的修正信号与高频异常，自主归纳并提炼为全新规则护栏与 Few-Shot，实现政策边界不确定问题改完即时生效，**驱动客服话术规则自进化生长。**

二、水陆客运场站边缘安检视觉分析与通行态势预警中台

倍施特科技 AI 应用研发工程师 2025.04 ~ 至今

> 针对水陆客运枢纽、码头航线（如宁波至象山航线票务通道）与景区进站通道人工看包易漏检、节假日闸机过客拥堵及违规滞留风险，将传统离线录像抽检升级为端边协同的秒级实时感知与态势预警中台。直连现场闸机、刷脸机、客流传感器与监控视频流，构建覆盖违禁品初筛、客流密度超限告警到现场处置分流的软硬一体化闭环体系。

推进节奏：① 站端物理设备与视频流协议对接 → ② 异步 VLM 视觉推理流水线 → ③ 客流过闸态势预警与状态机闭环

- **站端多源物理设备接入与点位管理：**深度对接现场安检闸机、刷脸通道、人流量传感器与网络监控摄像头，通过 TCP / HTTP 混合协议接入设备遥测流；构建物理设备点位映射与心跳保活机制，自研断线自愈重连策略，保障边缘硬件长周期高频运转不断流。
- **视频流关键帧自适应抽取与抽取优化：**基于 FFmpeg.wasm 与轻量级运动目标检测算子实现安检视频流自适应抽帧，过滤冗余静态背景帧，降低下游推理算力开销 60%；集成 OCR 文本识别提取行李牌与面单信息，结合图像元数据构建图文结构化输入通道。
- **Qwen2-VL 视觉大模型图文端到端推理：**集成 Qwen2-VL 视觉多模态大模型对包裹 X 光切片与过通道人员开展端到端推理，输出结构化违禁告警 JSON；将场站安检规范与危险物品级别抽象为可动态热更新的策略规则表，实现算法逻辑与安检业务解耦。
- **异步网关削峰与事务状态机高可用容错：**利用 Python asyncio 构建异步并发网关削峰；基于事务机制设计分析任务状态机，仅在全链路处理成功后提交终态，失败任务自动安全回滚并重新入队，结合僵尸任务扫描保障恶劣工况下不漏包、不丢任务。
- **客流拥堵态势预警与现场联动闭环：**基于站端通行数据建立人流密度建模，在节假日与高峰期实时计算过闸吞吐率；超阈值时自动触发红外预警、联动调度大屏警示与站端广播分流，复核数据沉淀回流评测集，形成「AI 初筛 → 现场处置 → 样本回流」闭环。

三、「云枢」客运数据治理中枢、客流时序预测与大模型统一网关

倍施特科技　**AI 应用研发工程师**　2025.06 ~ 至今

> 沉淀客服系统海量多轮对话、站端预警事件及售票流水等异构数据，构建「云枢」客运 AI 数据中台进行统一清洗与标准化治理。基于沉淀数据向文旅景区与客运枢纽提供节假日客流峰值与票务筹备预测服务，并纳管集团各业务线的大模型调用网关，使上层应用与底层基础设施彻底解耦。

· 阶段一：多源海量会话数据清洗脱敏与标准化数据工程

- **ETL 数据清洗脱敏流水线：**针对线上历史会话与站端多模态日志，设计时间窗口分句与正则/NER 多层脱敏流水线；剥离无效寒暄，脱敏手机号、身份证等敏感信息（准确率 98%），沉淀出万级高质量标准微调样本与 RAG 规程资产。
- **多格式训练数据导出与数据飞轮：**构建自动化数据导出管线，支持快速导出为 Alpaca / ShareGPT / JSONL 等多种通用训练格式；将线上评测发现的 Badcase 与人工修正样例自动归流标注池，驱动模型与检索底座持续迭代。

· 阶段二：客流态势分析、节假日运力预测与景区运营反哺

- **客流态势建模与运力票务预测：**整合线上售票意向、历史出行规律与现场过闸热度构建客流分析模型；在节假日高峰期前夕输出客流峰值预测报告，为文旅景区提供门票配额、接驳运力与餐饮商铺开设建议，形成运营数据反哺闭环。

· 阶段三：统一大模型网关治理、成本核算与运维态势大屏

- **虚拟密钥体系与多模型动态路由：**搭建集团统一大模型网关，纳管底层厂商真实 API Key，面向各业务线签发虚拟 Key 并挂载配额上限；基于任务复杂度实施动态路由，常规任务走轻量模型、复杂政策走垂直模型，替换基座业务层零感知。
- **全链路 Trace 追踪与运维监控智能体：**全链路透传 Trace ID，按业务线、场站精准统计 Prompt/Completion 的 Token 消耗与成本核算；配置滑动窗口接口限流与熔断护栏；构建运维监控 Agent，实时监测 Provider 延迟与错误率，联动大屏展示算力水位，月均 API 成本保持稳定受控。

四、管家婆辉煌小工单 MES 与「小辉智能体」工业排障

成都任我行　**AI 全栈开发工程师**　2024.03 ~ 2024.12

> 面向离散制造企业一线车间，负责管家婆辉煌小工单 MES 全栈迭代与机台设备数据采集打通；后期针对车间报障痛点从 0 到 1 打造「小辉智能体」，构建多 Agent 协同的车间自动化排障体系，车间排障响应时间从 4 小时压缩至 15 分钟，OnCall 咨询量下降 40%。

· 阶段一：MES 业务系统全栈迭代与底层机台数据通信打通

- **MES 工业业务系统全栈现代化重构：**负责生产工单派发、实时报工、批次溯源与开票业务数据流打通；主导老旧 **FineUI** 框架向 .NET Core Web API + Vue3 的平滑重构，优化 SQL Server / PostgreSQL 查询支撑车间多班次连续运转。
- **车间生产机台点位通信集成：**深入车间主导 MES 系统与产线 PLC、下位控制设备的数据通道对接，基于 Modbus / TCP 协议采集设备启停、加工计数与故障告警代码，建立物理点位与工单的映射，实现生产数据实时回传。

· 阶段二：多 Agent 协作排障工作流与工业专家 RAG 知识底座

- **多智能体协作排障工作流：**基于 Plan-Execute-Replan 与 ReAct 架构拆分对话理解、告警诊断与知识检索三大 Agent；借助 **PostgreSQL Checkpointer** 维护排障长会话状态，打通内部监控告警库与历史维保档案，实现长链路排障断点可恢复。
- **工业错误码专家知识库与现场闭环：**针对传感器异常报文与设备告警代码构建知识底座，采用父子块双层检索结合重排算法，规避切片碎片化导致的语义断层；打通报修派单链路，根因定位与排障手册秒级推送到产线工人移动端，大幅降低工程师重复 OnCall 咨询量。

荣誉认证 & 开源生态

- **工信部人才专业能力提升证书（高级）：**人工智能技术应用方向，涵盖大模型应用实施与软件工程开发。
- **GitHub 开源实践（Qiang-Maker）：**长期维护个人开源仓库，输出 Agent 运行时、Agent OS 协作工作台与线上产品。

教育背景

- **陕西机电职业技术学院**　**物联网应用技术**　2022.09 ~ 2025.07