童潇波

男 | 年龄：38岁 | 15968437118 | rotas.tong@gmail.com
15年工作经验 | 求职意向：大模型算法 | 期望城市：宁波

个人优势

多年软件工程经验,具备从 0 到 1 独立交付完整产品的全链路能力,技术覆盖前端、移动端、后端服务及 AI 系统工程。近年聚焦 AI 方向,已独立完成多个对话式 AI Agent 系统及多模态 AI 应用的架构设计与工程落地,深度掌握 LLM 集成、MCP (Model Context Protocol) 工具编排、Function Calling、Agentic Workflow、RAG (向量 + 全文 Hybrid Search)、多 provider LLM 抽象与 fallback chain、多模态(视觉理解 + 图像生成)、Prompt Engineering、Structured Output 与 LLM 成本/可靠性优化(Batch API、Embedding 缓存、指数退避、批 API 熔断)等核心技术。长期担任技术负责人角色,习惯在资源有限的条件下独立推进架构决策与产品交付,擅长将新技术与实际业务场景结合落地。GitHub 开源多个项目,涵盖多 LLM Agent 平台、AI 内容创作系统、PDF 智能翻译、自动化测试系统、Agent 资源录入与分享平台等;熟练运用 Claude Code、Codex 等 AI 辅助开发,显著提升交付效率。

工作经历

蓝卓数字科技有限公司　大模型算法　2018.08-2026.04

作为技术负责人,独立规划并交付公司工业互联网操作系统移动产品线,主导移动端架构设计与组件化拆分,搭建 Native 与 H5 / 小程序混合开发体系,落地 IM、音视频通话、AI 识别、工业外设、消息推送等核心能力。
基于已有 IM 体系,从零设计并落地对话式 AI Agent 工业生产系统,支撑告警处置、工单流转等场景的自然语言操作。完成 Agent 运行时、多 provider LLM 编排(Claude / Bedrock / Gemini / OpenAI 多协议适配 + fallback chain)、MCP (Model Context Protocol) 工具编排、Hybrid RAG、Function Calling、Prompt Engineering 等核心模块工程落地;实现 Agentic Workflow 全链路闭环(Reasoning → Planning → Confirmation → Action → Observation),构建 Guardrails、Structured Output、confirmation-gated 执行、用户级凭证隔离 等安全治理体系。

浙江甬润科技有限公司　Java　2017.03-2018.08

1. 负责公司多款社交娱乐类应用的开发，涵盖直播社交、实时游戏等业务场景
2. 独立承担狼人杀App安卓版本的全部开发工作，从架构设计到功能上线全程负责
3. 参与觅秀、花间等社交类App的核心功能模块开发与迭代

宁波普天通信技术有限公司　Java　2011.06-2017.02

1. 从零构建面向移动入库测试场景的自动化测试工具体系，服务于多家手机终端厂商的产品入库验证流程
2. 参与工信部安测云平台建设，为手机终端厂商提供在线化测试能力（对标腾讯优测），支撑终端检测业务的线上化转型
3. 主导商用终端测试系统 Android 前端与后台的研发工作，持续迭代近 3 年，保障系统稳定运行
4. 引入 Face++ 人脸识别技术落地智能门禁项目，实现刷脸开门的完整闭环
5. 在职 6 年间完成从测试脚本开发到 Android 应用开发、后台系统开发的技术成长，具备端到端的全栈交付能力

项目经历

TechPackAnnotator　项目负责人　2026.01-2026.04

1、设计 provider-agnostic LLM 抽象层(BatchedTranslatorBackend ABC + 工厂模式),基于 Anthropic-compatible 协议适配,支持 Claude / MiniMax / OpenAI / Gemini 等多 LLM 提供商插拔接入;LLM 切换 / A-B 对比 / 故障降级在架构层面解耦
2、设计 5 级混合翻译流水线(白名单 → 标识符正则 → 领域术语表 → 翻译记忆库 → LLM 兜底),将 LLM 调用率压到 ~15%,token

成本下降约 6 倍

3、设计 Token Masking + ID-keyed JSON 双向协议:翻译前将关键标识符 / 数字 / 单位替换为 __T0__ 占位符送入 LLM,输出后还原;LLM 以 {id: text} JSON 双向收发,通过 ID 而非位置匹配回原 bbox,从协议层杜绝 LLM 翻译丢失关键字段或输出乱序

4、实现 指数退避重试 + 批次级故障降级 + 速率限制感知并发 的高可用 LLM 调用层(ThreadPoolExecutor 多路并发批量调度),单文档 700+ 文本块端到端 P95 < 30 秒

5、实现 数据库驱动的翻译记忆库 (Translation Memory) + 双语术语表 (Glossary) 双层 LLM 结果缓存,同类文档二次处理 LLM 调用量再降 50%+,冷启动 30s → 热处理 < 5s

6、针对密排表格场景调优 PaddleOCR PP-OCRv5(ONNX Runtime)DBNet 检测参数,解决默认参数下相邻表格单元格被错误检测合并为单一 bbox 的问题;基于 Strategy + Factory 实现多 OCR 后端可插拔架构(Mock / PaddleOCR / PP-Structure / PP-OCRv5)

7、实现 细粒度异步流水线(渲染 → OCR → 翻译 → 标注 4 阶段进度上报)+ 增量 UI 解锁:渲染阶段完成即可打开编辑器,翻译标注在后台陆续就位,显著改善长文档场景体验

技术栈:Python · FastAPI · SQLAlchemy 2.0 · Alembic · PostgreSQL · Celery · Redis · MinIO · Anthropic SDK · PaddleOCR PP-OCRv5 · ONNX Runtime · React 18 · TypeScript · Vite · react-konva · Docker Compose

AI 驱动的多阶段智能内容创作平台　项目负责人　2026.01-2026.04

1、设计 provider-agnostic LLM 编排层,基于 Vercel AI SDK 同时支持 OpenAI-compatible 与 Anthropic Messages 双协议,实现 OpenAI / Claude / Kimi / MiniMax / Gemini 等 LLM 提供商插拔接入;Vision provider 与文本 provider 独立配置,支持成本优化的混合 provider 组合

2、构建 5 阶段 LLM Agent 流水线(任务理解 → RAG 召回 → 策略 → 文案 → 图像规划与生成),各阶段输出严格 JSON Schema 验证 + TypeScript 类型守卫,Agent 间结构化数据传递,参考来源全程可追溯

3、实现 混合 RAG 系统(Hybrid Search):pgvector 语义检索(text-embedding-3-small 1536 维)+ PostgreSQL 全文检索词汇召回,embedding provider 不可用时自动降级到 lexical-only,零生成中断;召回不足时级联放宽过滤条件(content_type → track → 全量),保证最低 5 篇候选

4、设计 结构化输出双模式 fallback:优先 generateObject + jsonSchema(temperature 0,确定性可复现);provider 不支持原生结构化输出时自动切换到文本 JSON 解析 + 部分 JSON 修复(parsePartialJson),从架构层适配不同 LLM 协议的 capability 差异

5、实现 零样本 / 检索增强双模式自适应:召回 top-k 相似度 < 0.6 自动降级到零样本生成 prompt,≥ 0.6 时注入 max-7 篇精选参考并通过 prompt 显式约束(禁止抄袭原句、要求抽象到底层逻辑),保证冷启动场景也能产出高质量内容

6、实现 多模态能力:视觉分析(单 prompt 完成 OCR + 视觉风格拆解 + 排版标签提取)+ 图像生成(OpenAI DALL-E + Google Vertex AI Imagen 双 vendor 适配,支持单批多候选与流式响应),含启发式图像规划 fallback

7、实现 流式 + 异步双轨:文案生成阶段 streamObject + SSE 流式输出,生成结果增量可见;BullMQ + Redis 三池工作队列(sample-analyze / sample-embed / image-generate)处理长任务,带 job 失败状态机与结构化日志(Pino)

技术栈:TypeScript · Next.js 16 · React 19 · PostgreSQL 16 · pgvector · Redis · BullMQ · Vercel AI SDK · OpenAI SDK · Anthropic SDK · Google Vertex AI · Pino · node-pg-migrate · Docker Compose

supLink AI Agent 系统　项目负责人　2025.10-2026.04

三层解耦的 Agentic 架构:IM Bot 接入层(XMPP)→ Agent 决策网关(Node.js / TypeScript)→ MCP 工具服务集群(Python FastMCP)。LLM 负责自然语言理解与执行计划生成,MCP 负责工具调用,confirmation gate 在协议层阻止 LLM 误调用敏感工具。面向企业 IM 场景实现自然语言驱动的告警查询 / 工作流编排 / 报表生成等任务自动化。

1、设计 provider-agnostic LLM 编排层,接入 Anthropic Claude / AWS Bedrock / Google Gemini / OpenAI 多家原生协议,架构层支持 MiniMax / Qwen / Doubao 等 provider 可插拔扩展;实现 3 层 fallback chain(primary → secondary → tertiary)+ 指数退避重试(429 / 5xx,最大 8s,3 次),LLM 单 provider 故障服务不中断

2、集成 MCP (Model Context Protocol) 工具编排层,通过 stdio transport 接入多个生产级 Python FastMCP 服务(告警 / 工作流 / 通知 / 报表),实现工具 JSON Schema 校验 + 参数绑定 + 结果标准化 完整执行管道,LLM 调用工具的可靠性与权限边界在协议层强约束

3、设计 Agent 多阶段决策状态机:Direct Answer → Plan Generation → Await User Confirmation → Formal Execution,confirmation-gated 执行(用户确认前 LLM 计划只生成不落地),从架构层杜绝 LLM 自动调用敏感工具的失控风险

4、实现 Hybrid RAG 系统:sqlite-vec 向量检索 + SQLite FTS5 全文检索双路并行(默认 0.7 / 0.3 权重融合)+ MMR reranking 多样性优化 + 时间衰减(默认 30 天半衰期);多 provider embedding 降级链路,所有 embedding provider 失败时自动降级到 FTS-only 模式,服务零中断

5、构建 LLM 成本优化体系:Embedding 批处理(OpenAI / Gemini / Voyage Batch API,50K 请求/批,成本下降 ~50%)+ SQLite 持久化 embedding 缓存(LRU,默认 100K 条目,典型场景调用复用率 60-80%)+ session 级 token / 成本细粒度追踪与配额控制

6、实现 生产级可靠性体系:批 API 熔断(连续 2 次失败自动降级到单请求模式)+ 多 provider fallback chain + fixed-window 速率限制器 + per-provider 配额追踪(5h / 7d 双窗口),从 LLM 抖动到 provider 长时不可用全场景兜底

7、实现 企业 IM Bot 接入层(基于 slixmpp XMPP):消息去重(10 分钟窗口 / 2000 条 LRU 缓存)+ 双向消息流(文本 / 结构化 UI 卡片 / 确认 stanza)+ Bot 与 Agent 网关解耦,支持 SSE 流式响应(agent lifecycle / tool / assistant 多事件流);Agent 工具集额外集成 Playwright 浏览器自动化 / Sharp 图像处理 / TTS 语音合成,均可被 LLM 编排调用

技术栈:TypeScript · Node.js 22 · Python 3.10+ · Express · Hono · slixmpp · FastMCP · SQLite · sqlite-vec · FTS5 · Anthropic SDK · AWS Bedrock SDK · Google Gemini SDK · OpenAI SDK · node-llama-cpp · LanceDB · Playwright · Sharp · TypeBox · pnpm monorepo · Docker

Suplink移动应用平台　项目负责人　2018.08-2026.04

作为唯一移动端负责人，独立完成 iOS / Android / HarmonyOS / 小程序四端技术架构设计与落地

设计小程序运行容器及 18 大类原生 API 开放层，支撑第三方工业应用接入；基于 XMPP 定制企业级 IM 引擎，集成 E2E 加密与音视频通话，集成 OCR、语音识别/合成、视觉检测、人脸识别等 AI 能力，完成 RFID、蓝牙、工业打印机等外设协议适配，落地工业巡检与数据采集场景

基于 Spring Boot + MyBatis 承担后端业务模块开发，负责 RESTful 接口设计、数据库表结构与索引优化

教育经历

浙江万里学院　本科　电子信息工程　2007-2011