

George Zhang

企业级 Agent 平台架构负责人

✉ sunionpeaks@gmail.com

🌐 中国

📅 10 年经验

概览

10 年大型平台与 AI 工程经验，历经阿里、商汤，现负责企业级 Agent 平台架构、产品化与团队交付。

服务规模

2k+ 租户 / 23w+ 用户

峰值请求

6w / day

团队范围

20 人产研 / 3 人 Agent 小组

Harness 覆盖

40 个仓库

经历

Agent 平台负责人

YC

2024-03 - 现在

负责跨行业 AI Agent 客服 SaaS 平台从架构演进到生产运营，覆盖 2k+ 租户、23w+ 用户、峰值 6w 请求/天；统筹 Runtime、工具与 Skill、RAG/Memory、质量评测、成本治理和团队交付。

主导模型驱动 Agent Runtime 2.0，基于 WorkflowAgent 建设 Agentic RAG、多轮工具调用、终止判断与多 Agent 并行编排；平均 7k+ tokens 对话负载下端到端约 7s，无工具场景约 1-2k tokens。

设计多 Agent 统一上下文机制，以稳定 prompt 前缀、动态上下文后置、职责节点隔离和阶段级 usage 审计提升 KV cache 复用率，并定位时延与 token 消耗来源。

建设动态 Skill 路由与工具按需挂载，按意图召回 SOP 和专属工具；多 Skill 命中时通过澄清与重新路由控制上下文长度、工具误调用和任务完成质量。

建立覆盖请求、执行阶段、工具调用与终态回传的观测和评测链路，以 correlation_id、trace、token usage、回放集和发布门禁定位质量、时延与成本问题。

建设未知问题聚类与主题发现平台，将每周约 1w+ 至 6w+ 条问题沉淀为 content gap；约 50% 聚类结果进入知识库/RAG/Skill 迭代，高频主题处理覆盖约 70%，自动解决率提升约 30%。

带领 3 人 Agent 小组推进 Runtime 服务化拆分与平台交付，统一网关鉴权、服务契约、容器构建和发布规范；推动 AI 开发 Harness 覆盖 40 个仓库，形成 Spec/EPIC、验收门禁、自检脚本和 PR 模板。

AIGC SaaS 架构师

SenseTime

2021-09 - 2024-05

负责 AI 绘画、模型训练与平台化产品的 Web 工程和微服务架构，参与生成式 AI SaaS 从 0 到 1 建设。

主导 miaohua 等 AIGC 产品的前后端架构、业务拆分、容器化部署和 DevOps 流程，支撑自研作画模型、LoRA 训练和第三方开源模型接入。

围绕 Stable Diffusion / PyTorch 推理链路建设模型参数格式化、prompt 优化、任务队列、模型服务分层、日志上报和监控能力。

设计身份认证、模型推理、模型文件存储、任务推送等微服务边界，并兼顾跨平台 Web / H5 / 小程序产品体验。

高级 Web 前端工程师

Alibaba Group - Taobao

2020-07 - 2021-08

负责淘宝直播、淘宝逛逛等高流量业务的跨端 Web 工程，重点在微前端编排、组件通信规范、SSR/边缘渲染和性能优化。

主导直播间 Web 互动层微前端方案，从 Weex 单组件模式演进到 H5 组件编排，统一布局、路由、事件总线 and 通信规范。

资深前端工程师

Alibaba Local Services

2017-10 - 2020-07

负责本地生活商业推广业务的广告投放、订单归因、跨平台 H5、小程序和商家后台工程，连接消费侧流量与商家经营数据。

设计基于 shell / iframe 的微前端架构，统一多业务子应用通信、路由定义、静态资源加载和部署方式。

沉淀 Vue / React 脚手架、跨平台 JsBridge、广告日志上报 SDK 和通用组件库，参与下单、广告结算、订单归因到商家经营分析的关键链路建设。

全栈开发工程师

Ningbo Senpu Rongxun

2016-05 - 2017-09

负责金融报表、Java Web 服务和前端迁移，完成 AngularJS 到 Vue 的客户端迁移和 Web UI 组件库建设。

基于 Spring Boot、MyBatis、AngularJS/Vue 建设金融报表数据汇总与业务管理系统。

能力

Agent 平台架构

Agent Runtime / Agentic RAG / Memory / Multi-agent Orchestration / Workflow Orchestration / Tool-use / Function Calling / MCP / Skills / Human-in-the-loop

生产治理与质量运营

Multi-tenant Isolation / Permission Model / Tool Registry / Audit Trail / Guardrails / Replay Evaluation / Golden Set / Trace Observability / Usage / Cost Governance

平台工程与企业集成

Python / FastAPI / TypeScript / Node.js / Rust / Java / Linux / PostgreSQL / Redis / Milvus / Docker / Kubernetes / Microservices / CI/CD / Alibaba Cloud / Google Cloud

模型与数据工程

Model APIs / Routing / Context Engineering / RAG Retrieval / Reranking / PyTorch / Stable Diffusion / Gemini / Data Feedback Loops / LLMOps

技术领导与工程治理

Platform Roadmap / Architecture Rules / Spec / EPIC Workflow / Release Gates / Cross-functional Delivery / AI Development Harness / Team Standards

教育

上海交通大学（网络教育）

本科 / 计算机科学与技术

2015-09 - 2018-07