

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

2025-03 - 现在

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

项目经验

WhatsApp SaaS 业务 Agent 平台

面向客服问答、业务查询、流程办理和人工转接，主导平台从规则编排演进为 LLM 驱动业务 Agent，将意图路由、SOP、知识、工具调用与结果回传沉淀为租户可配置能力；支撑客户问题自动解决率达到 70%+，平均 7k+ tokens 对话负载下端到端约 7s，token 消耗下降 60%。

企业级 Agentic RAG 与知识自进化

建设多租户 Agentic RAG，统一文档、QA、网站、附件、商品等知识接入、切分、embedding、召回与重排；以 unknown questions 聚类 and content gap 回流驱动知识库、RAG 与 Skill 迭代，约 50% 聚类结果进入生产优化，高频主题处理覆盖约 70%，自动解决率提升约 30%。

注意力引导推理 Agent 与复杂任务编排

基于 LangChain 有向无环图建设注意力引导推理 Agent，以 Guideline、Journey 和 ARQ 将复杂业务任务拆解为意图、规则、工具、参数与合规判断节点，管理分支依赖、终止条件和异常回退，形成可解释、可回放的多步骤执行链。

用户与会话 Memory 体系

建设会话级与用户级 Memory，统一事实提取、置信筛选、写入更新、召回注入和租户隔离；将历史结论、用户偏好与任务状态按需带入后续交互，减少重复确认并提升长对话和连续任务的一致性。

企业 AI 协作 CLI 与共享上下文平台

主导内部 AI 协作 CLI 与全员共享上下文机制，覆盖生产发布、运维自动化、case 排查、项目组织和工程自检；整合 Skill、MCP、CLI、仓库规范与 Spec/EPIC，将服务、环境和项目上下文沉淀为团队共享能力，推动 AI 开发 Harness 覆盖 40 个仓库。

AI 平台架构与全栈研发

自主创业

2024-05 - 2025-02

围绕 AI 数据与模型服务产品开展独立研发和项目交付，负责产品定义、系统架构、全栈实现、容器化部署与生产迭代。

项目经验

多模态数据标注与模型服务平台

从 0 到 1 打通人工标注、训练数据、模型产物、自动标注与质量校验闭环；以 React Canvas、Django、PostgreSQL 和 Redis 承载任务与数据链路，通过 Nuclio、Traefik 封装可插拔 Serverless 推理服务，完成音视频、图像任务的企业生产环境交付。

AIGC SaaS 架构师

商汤科技

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。

项目经验

妙画 AIGC SaaS 与模型服务链路

面向自研作画模型、LoRA 训练和第三方开源模型，主导前后端架构与服务拆分，将 Stable Diffusion / PyTorch 推理、任务队列、模型文件、结果后处理和日志监控组织为可扩展的异步生产链路，支撑 Web、H5 与小程序交付。

高级 Web 前端工程师

阿里巴巴集团

2017-10 - 2021-08

先后负责本地生活商业推广与淘宝新零售技术团队的高流量跨端工程，覆盖广告投放与归因、直播互动、微前端、SSR 和性能优化。

项目经验

淘宝直播跨端互动与微前端平台

主导直播间 Web 互动层从 Weex 单组件模式升级为 H5 微前端编排，统一布局、路由、事件总线和通信协议，并通过 SSR 同构渲染与分级资源策略将 FCP 稳定在 1.5s 以内，支撑高流量、弱网环境下的多业务组件交付。

全栈开发工程师

宁波森浦融讯

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。

项目经验

金融报表与业务管理平台

基于 Spring Boot、MyBatis 建设金融报表数据汇总与业务管理服务，完成 AngularJS 到 Vue 的渐进式迁移并沉淀共享 Web UI 组件，形成跨端平台架构与全栈交付的工程基础。

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