

George Zhang

Enterprise Agent Platform Architect

✉ sunionpeaks@gmail.com
🌐 China
📅 10 years

PROFILE

10 years across large-scale platforms and AI engineering at Alibaba and SenseTime, now leading enterprise Agent platform architecture, productization, and delivery.

Service scale

2k+ tenants / 230k+ users

Peak traffic

60k / day

Team scope

20-person product-engineering / 3-person Agent squad

Harness coverage

40 repositories

EXPERIENCE

Agent Platform Lead

YC

2025-03 - Present

Owned product architecture and production operations for a cross-industry WhatsApp business-Agent SaaS serving 2k+ tenants, 230k+ users, and 60k peak requests/day across RAG/Memory, tools and Skills, quality, cost, and team delivery.

Selected Projects

WhatsApp SaaS Business Agent Platform

Evolved WhatsApp customer service from rule flows to tenant-configurable LLM business Agents across intent routing, SOPs, knowledge, tools, callbacks, and handoff; delivered 70%+ auto-resolution, about 7s end-to-end latency at 7k+ tokens per conversation, and 60% lower token use.

Agentic RAG and Knowledge Flywheel

Built multi-tenant Agentic RAG for documents, QA, websites, attachments, and product data, with ingestion, retrieval, reranking, and an unknown-question/content-gap flywheel; 50% of clusters entered production optimization, high-frequency coverage reached 70%, and auto-resolution improved about 30%.

Attentive Reasoning for Complex Workflows

Built a LangChain DAG-based attentive reasoning Agent that decomposed workflows into intent, rule, tool, parameter, and compliance decisions, with explicit dependencies, termination, fallbacks, and replayability.

User and Session Memory

Built tenant-isolated session and user Memory for fact extraction, confidence filtering, write/update, and retrieval injection, carrying prior conclusions, preferences, and task state across interactions.

AI Collaboration CLI and Shared Context

Led an internal AI collaboration CLI and team-wide shared context for releases, operations automation, case diagnosis, project organization, and checks; integrated Skills, MCP, CLI, repository standards, and Spec/EPIC workflows across 40 repositories.

AI Platform Architecture & Full-stack Engineering

Independent Venture

2024-05 - 2025-02

Selected Projects

Multimodal Data Labeling and Model-serving Platform

Built a production multimodal data and model platform spanning annotation, model artifacts, auto-labeling, quality validation, and pluggable Nuclio/Traefik inference for audio, video, and image workflows.

AIGC SaaS Architect

SenseTime

2021-09 - 2024-05

Selected Projects

Miaohua AIGC SaaS and Model-serving Workflow

Led service architecture for proprietary image models, LoRA, and open-source integration, organizing Stable Diffusion/PyTorch inference, queues, model files, post-processing, and observability into asynchronous Web, H5, and mini-program delivery.

Senior Web Frontend Engineer

Alibaba Group

2017-10 - 2021-08

Selected Projects

Taobao Live Cross-platform Interaction and Micro-frontend Platform

Moved Taobao Live interaction from isolated Weex components to H5 micro-frontends with unified routing and communication; combined SSR and tiered resources to keep FCP below 1.5s under high traffic and weak networks.

Full-stack Developer

Ningbo Senpu Rongxun

2016-05 - 2017-09

Selected Projects

Financial Reporting and Operations Platform

Built Spring Boot/MyBatis financial operations services, migrated AngularJS to Vue, and established a shared Web UI library.

SKILLS

Agent Platform Architecture

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

Production Governance and Quality

Operations

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

Platform Engineering and Enterprise

Integration

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

Model and Data Engineering

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

Engineering Leadership and

Governance

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

EDUCATION

Shanghai Jiao Tong University (Online Education)

B.S. in Computer Science and Technology (Online Education)

2015-09 - 2018-07