

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	Peak traffic
2k+ tenants / 230k+ users	60k / day
Team scope	Harness coverage
20-person product-engineering / 3-person Agent squad	40 repositories

EXPERIENCE

Agent Platform Lead

YC

2024-03 - Present

Owned a cross-industry AI Agent customer-service SaaS platform from architecture evolution through production operations, covering 2k+ tenants, 230k+ users, and 60k peak requests/day across Runtime, tools and Skills, RAG/Memory, quality evaluation, cost governance, and team delivery.

Led model-driven Agent Runtime 2.0 with WorkflowAgent across Agentic RAG, multi-turn tool calls, termination decisions, and parallel multi-agent orchestration; held end-to-end latency around 7s under 7k+ average conversation-token load, with no-tool scenarios around 1-2k tokens.

Designed unified context for multi-agent execution, using stable prompt prefixes, deferred dynamic context, role-isolated nodes, and phase-level usage auditing to improve KV cache reuse and pinpoint latency and token drivers.

Built dynamic Skill routing and on-demand tool mounting, retrieving SOPs and dedicated tools by intent while resolving ambiguous multi-Skill matches through clarification and rerouting to balance context length, tool misuse, and task quality.

Established end-to-end observability and evaluation across requests, execution phases, tool calls, and terminal callbacks, using correlation IDs, traces, token usage, replay sets, and release gates to isolate quality, latency, and cost issues.

Built an unknown-question clustering and topic-discovery platform that turned about 10k+ to 60k+ weekly questions into content gaps; about 50% of clusters fed knowledge base/RAG/Skill iteration, high-frequency topic handling reached about 70%, and auto-resolution improved by about 30%.

Led a 3-person Agent squad through Runtime service decomposition and platform delivery, standardizing gateway authentication, service contracts, container builds, and release conventions; expanded the AI development harness across 40 repositories with Spec/EPIC workflows, acceptance gates, self-check scripts, and PR templates.

AIGC SaaS Architect

SenseTime

2021-09 - 2024-05

Owned web engineering and microservice architecture for AI image generation, model training, and platform products, contributing to 0-to-1 generative AI SaaS delivery.

Led frontend/backend architecture, domain decomposition, containerized deployment, and DevOps workflows for AIGC products such as miaohua, supporting proprietary image models, LoRA training, and open-source model integration.

Built model parameter formatting, prompt optimization, task queues, model-service layering, logging, and monitoring around Stable Diffusion / PyTorch inference workflows.

Designed service boundaries for identity, model inference, model-file storage, and task delivery while supporting cross-platform Web / H5 / mini-program product experience.

Senior Web Frontend Engineer

Alibaba Group - Taobao

2020-07 - 2021-08

Worked on high-traffic Taobao Live and Taobao content scenarios, focusing on micro-frontend orchestration, component communication contracts, SSR/edge rendering, and performance optimization.

Led the Taobao Live web interaction micro-frontend solution, evolving from isolated Weex components to H5 component orchestration with unified layout, routing, event bus, and communication contracts.

Senior Frontend Engineer

Alibaba Local Services

2017-10 - 2020-07

Owned ad delivery, order attribution, cross-platform H5, mini-program, and merchant-console engineering for Alibaba Local Services commercial promotion, connecting consumer traffic with merchant operating data.

Designed a shell / iframe based micro-frontend architecture to standardize sub-application communication, routing, static resource loading, and deployment.

Built Vue / React scaffolds, cross-platform JsBridge, ad logging SDK, and reusable components, contributing to key flows from ordering and ad settlement to attribution and merchant analysis.

Full-stack Developer

Ningbo Senpu Rongxun

2016-05 - 2017-09

Owned financial reporting, Java web services, and frontend migration from AngularJS to Vue with a shared Web UI component library.

Built financial-reporting aggregation and business management systems with Spring Boot, MyBatis, AngularJS, and Vue.

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