

靳泽旭 / Zexu Jin

Frontier Agent Harness Engineer · Agent Runtime / Tool Use / Context / State / Evals · Backend & AI Infrastructure

Tel: 18158522238 · Email: jackjin1997@gmail.com · GitHub: github.com/jackjin1997 · Site: jzxwiki.com · Education: B.S. Software Engineering

Summary

Backend engineer with nearly 6 years of production engineering experience and around 3 years focused on LLM applications and AI Agent systems. Led a 5-person backend team and built the engineering harness around agents: Agent Loop, Tool Use dispatch, Context construction, State / Checkpoint, Multi-Agent orchestration, MCP integration, observability, and evaluation loops.

In 2026, contributed multiple merged upstream PRs to LangChain / LangGraph / LangChainJS around agent runtime, checkpoint, streaming, and tool-use edge cases. Also briefly participated in DingTalk AI Agent direction, adding enterprise collaboration product context.

Keywords: Agent Harness, Agent Runtime, ReAct Agent Loop, Tool Calling, MCP, LangGraph Checkpoint, LangSmith, LLM-as-Judge, Context Engineering, Multi-Agent, Claude Code, Cursor, Python, TypeScript, FastAPI, Hono, Kubernetes, Redis, etcd.

Open Source Evidence

- **LangChain / LangGraph ecosystem:** 11 merged upstream PRs across langgraphjs, langchainjs, and langchain. Representative fixes: checkpoint backend consistency, special-channel writes, metadata filters, mixed Gemini built-in/function tools, streaming function_call chunks, and edge-case crash prevention.
- **Representative PRs:** langgraphjs #2516 / #2517 / #2518 / #2208 / #2186, langchainjs #10676, langchain #36108 / #34943.
- **Other merged PRs:** OpenViking #800 Kubernetes Helm chart, openclaw #47465 markdown parsing fix, Antigravity Awesome Skills #61 / #69, Cosmos SDK #25863.

Experience

钉钉 AI Agent 方向 · 短期工程经历

2026.06 · ~2 weeks · Enterprise collaboration AI Agent / Wukong-related direction

- Quickly learned business boundaries, permission constraints, and engineering collaboration model around enterprise Agent / CLI / toolchain scenarios.
- Listed as a short supplementary experience; core project depth remains from long-term Agent Harness delivery and open-source contributions.

Open Source / Agent Infrastructure · Upstream Contributor

2026.03 — 2026.06 · LangChain / LangGraph / LangChainJS / OpenViking

- Submitted and merged agent-runtime fixes around checkpoint backend consistency, special-channel writes, streaming function calls, Gemini mixed tools, and deployment infrastructure.
- Used real framework failure modes to build judgment around agent state, tool invocation protocols, context persistence, and reliable developer experience.

跨越星空科技有限公司 · Backend Engineer → Backend Tech Lead

2022.09 — 2026.02 · AI Agent and intelligent automation products · Led 5-person backend team

- Owned architecture, design review, and implementation direction for AI Agent product runtime and backend infrastructure.
- Led IMEANAI-Coyage Agent Harness from 0 to 1, including multi-agent orchestration, ReAct loop, dynamic tool selection, MCP tool integration, context/state handling, streaming inference, observability, and reliability mechanisms.
- Designed an AI online service framework in TypeScript + Hono as the shared runtime foundation for AI microservices, supporting typed actions, streaming RPC, Zod validation, etcd-based leader election, dynamic config, and typed SDK generation.
- Promoted AI-native engineering workflow with Claude Code / Cursor, human-controlled architecture and quality gates, Issue to implementation to PR review workflow, and repository-level CLAUDE.md / rule files.
- Earlier led backend delivery for an AI-native workflow orchestration platform similar to Dify / N8n, serving banking and ecommerce B2B customers.

FTL Group / 法翔物流 · Full-stack Developer

2020.07 — 2022.07 · French ecommerce logistics SaaS for European B2B market

- Worked on esendeo logistics SaaS: SQL and reconciliation optimization, AWS Lambda asynchronous PDF rendering, pricing engine refactor with table splitting and Redis caching.
- Built esendeo-shopify from 0 to 1 with Next.js and Koa, connecting Shopify order sync to logistics status updates.
- Built internal UI / ORM components and Puppeteer / Playwright crawlers for address and customs code workflows.

Selected Projects

IMEANAI-Coyage · Multi-Agent Travel Assistant / Agent Harness

2024.01 — 2026.02 · Backend Tech Lead / Agent Architect · Python, FastAPI, LangChain, LangGraph, LangSmith, MCP, Claude, OpenAI, Bedrock, Qwen, Pinecone, Docker, Kubernetes

- **Problem:** Natural-language travel planning requires reliable tool use and real-time validation; pure rules cannot handle complex intent, while pure LLM output can hallucinate unavailable itinerary items.
- **Agent orchestration:** Built Supervisor + 6 expert agents for hotel, flight, restaurant, attraction, itinerary, and currency domains. Each sub-agent runs a ReAct-style loop with reasoning, tool calling, observation, and self-verification.
- **Tool reliability:** Dynamic Tool Selection injects only relevant tool descriptions; MCP standardizes tool registration; Zod schema validation, timeout handling, fallback, and degradation protect the main path.
- **State and context:** LangGraph Checkpoint and Time Travel support long conversation recovery and node-level rollback; token-threshold summarization controls context growth.
- **Evaluation and observability:** LangSmith trace across Tool Call / Chain / LLM levels, custom event dispatch tagging, and offline LangSmith Dataset + LLM-as-Judge loop for prompt, harness, and tool-description iteration.
- **Outcome:** Production product with improved reliability through multi-LLM fallback, tool circuit breaking, Redis buffering, prompt cache, streaming UX, and declarative MCP tool onboarding; TTFB controlled within 10 seconds for complex intents.

AI Online Service Framework · imean-service-engine

2024.01 — 2026.02 · Solo design / core developer · TypeScript, Hono, Zod, Etdcd3, Redis, EJSON, Winston, LRU-Cache

- Built a lightweight, type-safe, extensible microservice framework for Agent and AI services, with decorator-based APIs, typed actions, RPC + AsyncIterator streaming, idempotency markers, and generated typed clients.
- Implemented etcd-based method-level leader election for distributed scheduled tasks, including agent evaluation regression jobs and vector index refresh jobs.
- Supported dynamic configuration hot updates for model, temperature, and prompt template parameters without redeployment.

AI-native Workflow Orchestration Platform

2022.09 — 2024.01 · Backend execution → Tech Lead · TypeScript, GraphQL, Prisma, React, Redis, MongoDB, RabbitMQ

- Built a B2B workflow orchestration platform similar to Dify / N8n, integrating LLM calls, AI decision routing, and vector retrieval into the execution engine.
- Implemented AtomFlow + CombinedFlow, nested / parallel workflow execution, multi-tenant RBAC, version rollback, JSON import/export, and audit tracking.

Skills

Agent Harness: ReAct Agent Loop, Tool Use, dynamic tool selection, MCP, Multi-Agent / Supervisor, Context Engineering, LangGraph Checkpoint, Time Travel, Agent State, Skills.

LLM / Evaluation: Claude, OpenAI, AWS Bedrock, Qwen, streaming LLM APIs, Prompt Cache, multi-model routing, LangSmith, LLM-as-Judge, trace-based failure attribution, Multi-LLM Fallback.

Backend / Infrastructure: TypeScript, Python, Go, FastAPI, Hono, GraphQL, NestJS, microservices, etcd, Redis, RabbitMQ, PostgreSQL, MongoDB, Docker, Kubernetes, GitLab CI, Linux performance analysis.

AI-assisted Development: Claude Code, Cursor, repository rule files, CLAUDE.md, Issue to implementation to PR review workflow.

Education & Certification

浙江大学城市学院 · Software Engineering · Bachelor · 2015 — 2019

Software Designer Certificate (Intermediate)