

Huizhirong (Jason) Guo

Jason.ghzr@gmail.com | +1 669-214-8407 | plonguo.com | linkedin.com/in/jasonguo1104 | github.com/PlonGuo

EDUCATION

Northeastern University — Khoury College of Computer Sciences

M.S. in Computer Science

Oakland, CA

Expected May 2027

Santa Clara University — College of Arts and Sciences

B.S. in Computer Science

Santa Clara, CA

Jun 2024

EXPERIENCE

Synthoid.ai — Embodied AI

AI Agent Algorithms Intern

Wuxi, China

Sep 2026 – Present

- Develop on-robot agent architecture and core algorithms; integrate agents with motion control and visual perception modules alongside cross-functional teams.

Bosch

AI Agent Engineer Intern

Wuxi, China

May 2026 – Aug 2026

- Built Dify email workflows for **100+ daily active users**, supporting batch summaries, multi-turn reply drafting, per-factory tenant routing and cost isolation, lightweight-model intent routing, and automatic fallbacks.
- Evaluated 30 real emails: replacing agent orchestration with workflows reduced P95 latency from **52s to 18s**; smaller models cut latency by another **50%+ on 70% of samples** without observed quality loss.
- Enabled per-user usage tracking without database access using parallel instrumentation and platform deduplication.
- Built a scheduling POC with five MySQL tables modeling material, fleet, and shift constraints; exposed queries through MCP and authored a deepagents skill for conversational scheduling.

Next Play Games

Software Engineer Intern

San Jose, CA

Sep 2025 – Dec 2025

- Led full-stack B2B SaaS development; prevented schema drift with versioned migrations, PR/post-merge CI validation, dev/production deployment, and rollback procedures.
- Used direct Supabase access with RLS and PostgreSQL RPC to reduce P95 latency from **~420ms to ~170ms**; replaced one-second mobile polling with Realtime subscriptions, cutting daily requests per client from **~29,000 to dozens**.
- Standardized AI-assisted development for 10+ teammates and onboarded three members; added AI PR reviews and automatic CI-failure issue creation through GitHub Actions.
- Built OpenAI-powered coaching and an automated fine-tuning pipeline: **200+ training examples and five iterations** without manual intervention.

Geshang Fuxin

Backend Engineer Intern

Beijing, China

Jun 2025 – Aug 2025

- Maintained and extended five RESTful APIs in a seven-year-old Flask/MongoDB codebase with backward compatibility, zero downtime, and no regressions.
- Replaced full-collection queries with pagination, reducing response times from **300–500ms to ~200ms**.

PROJECTS

KnowHive — Local-first RAG + Agent Knowledge Base Desktop App

[\[Github Repo\]](#)

TypeScript/Bun · Vercel AI SDK v7 · Ollama + DeepSeek · Langfuse

- Built a retrieval-augmented generation (RAG) pipeline with a unified document representation, document-specific chunking, and parent-child chunks. Combined **vector and FTS5 keyword search through reciprocal rank fusion**, followed by **cross-encoder reranking and parent-context expansion**.
- Improved RAGAS context precision from **0.788 to 0.862** in A/B evaluation before enabling the new pipeline by default; fixed a silent Chinese keyword-search failure caused by FTS5 tokenization, increasing recall by 0.054.
- Added **persistent memory** that distilled prior conversations into semantic, preference, and episodic memories; improved accuracy from **0% to 100%** on **six evaluation questions** whose answers appeared only in past conversations.
- Increased correct refusals on unanswerable questions from **32% to 91%**, with a **4% false-refusal rate**, using reranking-score thresholds. Isolated untrusted retrieved content through spotlighting, **reducing prompt-injection attack success from 40% to 27%**; a custom red-team test suite caught and helped resolve a later regression before merge.

Go Chat Room — Real-time Communication Server

[\[Github Repo\]](#)

Go · Gin · Redis · WebSocket / WebRTC

- Built a backend supporting **direct messaging, group chat, and peer-to-peer video calls** using dedicated read/write goroutines, buffered channels for non-blocking messaging, and **separate backpressure policies for chat and signaling**.
- Implemented **stateless JWT authentication without database lookups** on business endpoints, Redis-backed session management, and **custom WebRTC signaling** with ICE candidate queuing and recovery for signaling and media connections.
- Identified **synchronous persistence in the hub event loop** as a throughput bottleneck and **proposed asynchronous persistence and hub sharding**.

TECHNICAL SKILLS

Languages & Data: Python, TypeScript/JavaScript, Go, SQL, PostgreSQL, MySQL, MongoDB, Redis, SQLite

Frameworks: Next.js, React Native, FastAPI, Flask, Gin, Vercel AI SDK, LangChain, LangGraph, deepagents

AI / LLM & DevOps: RAG, agent tools, MCP, fine-tuning, RAGAS evals, Langfuse, Dify, Docker, GitHub Actions, Supabase