

Roy Zhao

(217) 991-2581 | royzhao2004@outlook.com | royzhao.dev | linkedin.com/in/jiaqi-roy-zhao | github.com/JiaqiZhao2004

Summary

Backend and platform-focused software engineer building APIs, database-backed applications, cloud infrastructure, and applied AI systems, with experience in deployment, production debugging, observability, and LLM performance optimization.

Education

University of Illinois at Urbana-Champaign (UIUC)

Aug 2023 -- May 2027

BS in Computer Science, GPA: 3.9 / 4.0

Relevant Coursework: System Programming, Database Systems, Cloud Computing Applications, ML Systems, IoT

Experience

CCC Intelligent Solutions

Chicago, IL

AI Enablement Intern

Jun 2025 -- Present

- Cut an internal agent's per-trace token cost **76%** (\$0.88 to \$0.21) and end-to-end latency **39%** (1m51s to 1m8s) by parallelizing tool calls, adding prompt caching, and pruning tool outputs; validated via Langfuse trace instrumentation
- Triaged deployment and runtime failures in the department's **production EKS** environment via Rancher and AWS Console, inspecting pod state, logs, and cluster events
- Promoted changes across multiple internal applications from development into production within the department using an established GitHub Actions CI/CD workflow
- Led PoC evaluations of LLM observability platforms, identified workflow gaps blocking complete evaluation cycles, and delivered recommendations that informed the department's tooling strategy

Association for Computing Machinery (ACM), UIUC

Champaign, IL

Infrastructure Team Member

Feb 2025 -- May 2025

- Shipped calendar UI formatting updates for ACM UIUC's production website serving 1,000+ members while contributing through code reviews, sprint planning, and a GitHub Actions CI/CD workflow

Projects

Lantern -- React, TypeScript, FastAPI, PostgreSQL, Plaid, Firebase, OpenAI

- Built and deployed a household finance application with Firebase auth, real and sandboxed Plaid account linking, webhook-driven transaction syncing, and PostgreSQL-backed dashboards
- Operated it on a Linux server with Docker Compose and Cloudflare tunnel ingress, using Grafana to monitor daily Plaid sync jobs and S3 backup freshness
- Detected stalled Plaid webhooks through Grafana, added production logging, and deployed a CloudFront Terraform fix that restored transaction syncing
- Implemented an AI-assisted Named Query workflow for recurring spending analysis, translating natural-language questions into household-scoped, read-only SQL and reusable dashboard views

Trace -- Vue, TypeScript, FastAPI, Lambda, API Gateway, DynamoDB, Cognito, CloudFront

- Operated a serverless time observability app in daily use for 5+ months, accumulating 5,300+ DynamoDB records through a Vue frontend and FastAPI Lambda backend
- Migrated from SQLite to DynamoDB for cost-efficient hosting, then used an expand-and-contract migration to transform thousands of existing records into a user-scoped key model without data loss
- Integrated Amazon Cognito with Google sign-in plus a shared **Try as guest** mode that lets visitors explore public tracking data while blocking profile/category mutations

Technical Skills

Languages: Python, TypeScript/JavaScript, SQL, C/C++

Backend/Data: FastAPI, REST APIs, PostgreSQL, DynamoDB

Cloud/Infrastructure: AWS (Lambda, API Gateway, S3, CloudFront, EKS), Terraform, Docker, Kubernetes, GitHub Actions

AI/LLM: LLM evaluation, Langfuse, LangGraph, Claude Agent SDK, CrewAI

Frontend: React, Vue