Jin **Tan Ruan**

Software Developer · Prototyping Engineer

(+1) 347-951-8537 | jtanruan@gmail.com | www.ztanruan.com | ztanruan | ztanruan

Recruitment Team *April 15, 2025*

New York, United States

About Me

I'm Jin Tan Ruan, a Senior Generative AI Engineer based in New York, with a focus on building and deploying production-ready applications powered by large language models (LLMs). My expertise spans agentic AI systems, fine-tuning foundation models, and developing scalable, enterprise-grade AI/ML solutions.

Currently, I serve as a Generative AI Developer on the Global Prototyping team at Amazon Web Services (AWS). In this role, I research, design, and develop next-generation AI applications that bring artificial general intelligence (AGI)-like capabilities into real-world enterprise workflows. I work closely with AI/ML scientists, solution architects, and engineers to build robust, scalable generative AI systems that solve complex business problems across multiple industries.

What I've Built

Over the past few years, I've designed and delivered production-ready AI solutions spanning multimodal agentic systems, edge AI, and scalable data infrastructure.

- **Agentic AI Chatbot – Multimodal Conversational Agent:** Built a generative AI chatbot using Amazon Bedrock's multi-agent orchestration to perform real-time web search, document analysis, and creative media generation. Integrated OCR, LLMs, and image/video generation to support enterprise use cases such as financial research, legal document processing, and customer support.
- **Edge AI with Jetson Nano – Real-Time Computer Vision:** Deployed YOLOv8 on Nvidia Jetson Nano for low-latency object detection at the edge. This solution enables real-time analytics for applications like smart surveillance, robotics, manufacturing quality control, and diagnostic imaging eliminating the need for cloud round trips.
- **SDAS – Scalable Data Archiving Solution:** Developed SDAS (Simple Database Archival Solution), a serverless, long-term data retention framework using AWS Step Functions, Glue, and Athena. Enabled secure archival of legacy databases to Amazon S3, cutting costs and supporting compliance in industries like healthcare and finance.

Why Me?

As a Senior Generative AI Engineer at AWS, I've designed, built, and shipped production-ready GenAI solutions including Retrieval Augmented Generation (RAG) pipelines, agentic systems, and multimodal applications. My core strength lies in blending deep knowledge of LLMs with strong software engineering fundamentals to deliver scalable, cloud-native architectures.

I've contributed to open-source projects, authored technical blogs on fine-tuning and prompt engineering, and built prototypes that have accelerated enterprise GenAI adoption. I thrive in environments where I can bridge the gap between research and engineering translating cutting-edge models into real-world, impactful AI experiences.

Sincerely,

Jin Tan Ruan

Jin **Tan Ruan**

Software Developer · Prototyping Engineer

(+1) 347-951-8537 | jtanruan@gmail.com | www.ztanruan.com | ztanruan | ztanruan

Education

Syracuse University - School of Engineering and Computer Science Syracuse, NY *Syracuse, NY*

M.S. In Computer Science And Engineering | GPA: 3.7 *Jan. 2020 - May. 2021*

- Blockchain & Cryptocurrencies, Computer Architecture, Deep learning, iOS Development, Cryptography, and Artificial Intelligence.

Syracuse University - School of Engineering and Computer Science Syracuse, NY *Syracuse, NY*

B.S. In Computer Science And Engineering | GPA: 3.7 *Aug. 2016 - Dec. 2019*

- Joined a selected group of promising students to the Concord Dawn - U.S Air Force Cybersecurity Program.
- Data structures and algorithms, Computer systems, Differential equations, Linear algebra II, Android Development.

Skills & Certifications

Certifications	AWS Certified Developer, AWS Certified Machine Learning, AWS Certified Solutions Architect Professional
Programming	Node.js, Python, Java, C++, LaTeX, XML, Haskell, Swift, SwiftUI, Objective C, Kotlin, Typescript
Back-end	REST API, MongoDB, MySQL, OpenSearch, PineCone, DynamoDB, Aurora, Oracle, S3, Kendra
Front-end	React, HTML5, PHP, JavaScript, JQuery, CSS, Bootstrap, Gatsby, Materia UI, Cloudscape, Vite, Next.js
DevOps	Amazon Web Services, Microsoft Azure, Google Cloud, Docker, Kubernetes, Terraform, CDK, Jenkins
Languages	Spanish (Native), English, Chinese (Native)

Work Experience

Amazon Web Services *New York City, NY*

Senior GenAI Prototyping Developer *Jan. 2022 - Present*

- Spearheaded the rapid prototyping and development of generative AI solutions across healthcare, automotive, and media sectors leveraging RAG, LLMs, and agent-based architectures to address real-world enterprise challenges.
- Played a key role in launching Amazon Bedrock agent capabilities, contributing to API schema design, simplifying developer workflows, and shaping the product roadmap through hands-on implementation.
- Designed an internal agentic catalog to orchestrate multi-agent collaboration for high-impact use cases such as drug discovery (HCLS) and financial market analysis (FSI), demonstrating advanced coordination and reasoning across autonomous agents.
- Published six open-source generative AI solutions, including RAG pipelines, image generation tools for marketing, a speech-to-speech agentic chatbot, and a multimodal chatbot powered by a multi-agent orchestrator.
- Authored four whitepapers and technical blogs on prompt engineering with LLaMA 4 and fine-tuning techniques using Low-Rank Adaptation (LoRA).

Deloitte Consulting LLP *New York City, NY*

Software Engineer | Full-Stack Java Developer *May. 2021 - Jan. 2022*

- Led the end-to-end design and development of enterprise-grade microservices using Amazon Web Services (AWS), Oracle Database, and JavaScript for a global leader in renal disease treatment, supporting over 190,000 patients across 2,400+ facilities nationwide.
- Designed and implemented scalable system APIs and background processing services to manage over 450 million messages and requests containing sensitive PHI and PII data, ensuring high availability and compliance.
- Provisioned and maintained a hybrid infrastructure (AWS + on-premises) using Infrastructure as Code (IaC) with AWS CloudFormation and the Cloud Development Kit (CDK), contributing to an estimated $8.3 billion in annual capital expenditure (CAPEX) savings.

Projects & Achievements

Multimodal Conversational Agent *New York City, NY*

Agentic AI Chatbot *Jan. 2025 - April. 2025*

- Built a generative AI chatbot using Amazon Bedrock's multi-agent orchestration for real-time web search, document processing, and creative media generation.
- Designed a multi-agent catalog spanning domains such as drug discovery, clinical trials, financial market analysis, and compliance workflows enabling intelligent collaboration between specialized agents.
- Demonstrated agent coordination, task routing, and reasoning across tools and modalities to support enterprise decision making.