Tiansi Gu

vitinia0425@gmail.com | linkedin.com/in/tiansi-gu-51796b267 | github.com/TiansiGu

Education

University of San Francisco — San Francisco, CA
Master of Science in Computer Science (GPA 4.0/4.0) — *Aug 2023 – May 2026*

Renmin University of China — Beijing, China
Bachelor in Labor Relations — *Sep 2017 – Jun 2021*

Experience

Amazon Web Services - Healthcare AI — May 2025 – Aug 2025
Software Development Engineer Intern — *Seattle, WA*

- Created an LLM-as-a-Judge evaluation framework for HealthScribe's clinical note summarization using Amazon Bedrock, which reduced false alarms by 57%, accelerating model release cycles on healthcare service platforms.
- Collaborated with data scientists to architect and build a extensible evaluation mechanism combining Bedrock's built-in prompts, custom prompts, and rule-based analysis with CloudWatch metric emission.
- Built scalable infrastructure via ECS Fargate to support large datasets with day-long execution windows; Configured KMS-encrypted secure S3 bucket storage and fine-grained access-controlled IAM roles using AWS CDK.
- Integrated and deployed the evaluations into integration/canary tests and runtime workflows in Java and Kotlin.

Amazon Web Services - Transcribe — May 2024 – Aug 2024
Software Development Engineer Intern — *Seattle, WA*

- Architected and implemented per-customer throttle limit controls for AWS HealthScribe, preventing LLM model overload from excessive clinical audio summarization requests and reducing latency by 64% during system scaling.
- Engineered secure and efficient CRUD APIs in Java to interact with DynamoDB to track customer job statuses.
- Implemented Lambda functions triggered by EventBridge to enforce limits, initiate NLP workflows in parallel, and ensure data consistency in a high-throughput environment.
- Achieved 97% unit test coverage and updated integration tests for the service change, supporting CI/CD pipelines.

Projects

SnapLogic's Agent Continuations with Platform MCPs - LLM System, AI Agents — Aug 2025 – Present

- Collaborating to deliver resumable AI agents in Python with human-in-the-loop approvals, LiteLLM multi-provider routing, streaming + retries, and OpenTelemetry tracing for performance visibility (team of three).
- Creating MCP servers to interact with SnapLogic APIs and MongoDB databases, enabling RBAC-controlled workflow diagnoses, asset navigations, and pipeline creation + validation through natural language.

NLP-Driven Product Recommendation System - Recommendation System — Feb 2025 – Apr 2025

- Built a semantic product recommender using Amazon Review Data by encoding review texts with Sentence-BERT embeddings and constructing a graph-based similarity network on PyTorch to identify related products.
- Enhanced recommendation diversity via Semantic Lexical Diversity (SLD) metrics from reviews, improving semantic coverage and variety over baseline embedding-only methods.

JobFit Checker - Fullstack Development, Infrastructure — Oct 2024 – Apr 2025

- Developed a web application with React frontend, Spring Boot backend, and PostgreSQL database that tells users if they qualify for a job and gives specific skill improvement suggestions.
- Leveraged HTML, CSS, and JavaScript to build 10+ reusable React.Js components. Built an event-driven workflow that processes S3 events of resume uploads via AWS SQS and extracts structured data utilizing LangChain4J.
- Developed a CI/CD operational pipeline with four promotion stages, enabling QA and blue-green deployment to production using Docker Containerization, Bash Scripts, Github Actions, ECR, and Elastic Kubernetes Service.

Technical Skills

Languages: Java, C++, C, Python, JavaScript, Typescript, Kotlin, SQL, HTML/CSS
Frameworks/Databases: Linux, Unix, React, ReactNative, REST APIs, CloudWatch, CloudFormation, Spring, Django, FastAPI, Maven, Git, JUnit, Jest, TestNG, MySQL, PostgreSQL, MongoDB, DynamoDB, Redis
AI/LLM: Machine Learning, PyTorch, Scikit-learn, Amazon Bedrock, OpenAI SDK, AI Agents
Cloud/DevOps: Google Cloud Platform, AWS, Terraform, Kubernetes, Nginx, Github Actions, Ansible
Latency Sensitive Service, Data Structures & Algorithms, Distributed System, Machine Learning, Parallel Programming