Yunquan Zhang

Greater Seattle Area | San Francisco Bay Area | yz2793@cornell.edu | (510)703-7879

Education

Cornell University 08/2022 – 12/2023

Master of Engineering, Electrical and Computer Engineering (GPA: 3.61/4.0)

South China University of Technology 09/2017 – 06/2022

Bachelor of Engineering in Information Engineering (GPA: 3.70/4.0)

Skills

Languages: Java, C, C++, Python, HTML/CSS, SQL, JavaScript/TypeScript, Kotlin

Utilities: Linux, Git, Docker, Nginx, Message Queue(Kafka, RabbitMQ), Redis, CI/CD, Github Actions, Kubernetes

Framework: Spring Boot, Django, FastAPI, Flask; React.js, Vue.js, Node.js, Next.js, Tailwind CSS, JQuery

AI Tools: TensorFlow, PyTorch, Keras, Scikit-learn, OpenAI API, LangChain

Cloud Platforms: AWS (Lambda, Bedrock, Glue, CDK, CloudWatch, EC2, S3, RDS), Azure (Static Web App, AKS)

Database Management Systems: MySQL, PostgreSQL, MongoDB, SQLite, Pinecone

Work Experience

Amazon, Payments Services 04/2025 – Now

Software Dev Engineer *Seattle, WA*

- **CORA AI Agent Chatbot**
  - Developed an **AI Agent** CORA 3o (Cost Reporting & Analytics) to automate email alerting and aggregate cost metrics (IMR, DIG and credit scores); used by **67 teams** in Payment Services.
  - Built a chatbot frontend within the **Harmony Console** using **React.js**, integrated with **AWS Bedrock** via **Lambda** functions, leveraging native **S3** knowledge base support to enable context-aware responses to user.
- **Chase Credit Card Auto-Pay API Development**
  - Implemented in-app Auto-Pay features for **Amazon–Chase Co-Branded credit cards** by designing recurring-pay **APIs** with secure authentication; now used by **millions of** Amazon Prime credit card holders.

CoScribe AI 08/2023 – 01/2025

Full-Stack Developer *Framingham, MA*

- **AI-Driven Multimodal Summarization & RAG System**
  - Built a multimodal summarization system to process videos, images, audio, and text into unified summaries, enhancing accessibility and usability across diverse media formats.
  - Developed a multimodal **Retrieval-Augmented Generation (RAG) system** using **LangChain, GPT-4, and Pinecone**, applying prompt engineering like **Chain of Thought (CoT)**, driving a **15% DAU increase**.
  - Designed an API backend with **Python FastAPI**, integrating **RabbitMQ** for asynchronous task handling, achieving **3GB/min**throughput and supporting thousands of concurrent requests per second.
  - Designed a hybrid storage solution with **MongoDB** for metadata, **Azure Blob Storage** for media, and **Pinecone** for semantic search, ensuring scalability and efficient retrieval for high-demand systems.
- Implemented a comprehensive **CI/CD pipeline** using **GitHub Actions**, automating **Docker** image builds and containerized deployments to **Azure Kubernetes Service (AKS)** with integration of **GitHub Container Registry (GHCR)**, reducing deployment time by **30%** and enabling auto-scaling capabilities.
- Developed web applications using **React.js** and cross-platform mobile apps with **React Native**, leveraging **TypeScript** for type-safe development and **Redux** for state management, now serving **5000+ users**.

Roamer AI 06/2023 – 08/2023

Software Developer (Intern) *Remote*

- Developed an advanced apartment search web app enabling seamless text and image-based queries by integrating cutting-edge AI models, improving user experience through innovative AI-powered solutions.
- Developed a full-stack solution with **Python Django** and **Vue.js**, leveraging Django **ORM** integrated with **PostgreSQL** to design and manage database schemas and implement efficient query handling. Deployed the application on **AWS EC2**, serving **500+** daily active users with 99.9% uptime.
- Scraped apartment data using **Python**, **Selenium**, and **Beautiful Soup**, storing the processed dataset on **Amazon S3** for scalability. Fine-tuned **OpenAI's CLIP model** on **AWS EC2** using **PyTorch**, achieving a 30% improvement in query accuracy through model optimization and fine-tuning.