

Songyan Zhao

linkedin.com/in/songyan-silas-zhao/

github.com/SilasZhao

✉ zhaosongyan7@gmail.com

☎ +1-5075811895

🌐 Google Scholar

SUMMARY

Seeking Machine Learning Engineer roles from 11/2025. Greencard holder. Full-stack developer with expertise in AI Infrastructure, machine learning, large language models, recommendation system, and human-AI interaction. First-author publications at top NLP conferences. Skilled in scalable web and backend development (Java Spring Boot, Python Flask, etc.).

EDUCATION

•University of California, Los Angeles	Sept. 2023 - June. 2025
M.S. Computer Science, advisor: Prof. Nanyun Peng, GPA: 3.88/4.0	Los Angeles, CA
•Carleton College	Sept. 2019 - Jun. 2023
B.A. Computer Science and Mathematics (double-majors); GPA: 3.86/4.0, Cum laude	Northfield, MN

TECHNICAL SKILLS

Languages: Python, Java, C, C++/STL, SQL, HTML, CSS, JavaScript, Go, Scheme, TypeScript
Frameworks: Spring Boot, Flask, Node.js, Vue.js, Pytorch, Angular, Node.js, .Net, JQuery, TensorFlow, Amazon AWS
Databases/Tools: MySQL, PostgreSQL, MongoDB, Core Data, Redis, Docker, Dplr, Gradle, Maven, gdb
Knowledge: OOP/OOD, Design patterns (MVC, Factory), Large language model, Machine learning, Distributed System, Relational databases, RAG.

WORK EXPERIENCE

•XPOWER Manufacture Inc.: Amazon Bedrock based AI agent	June. 2025 - Nov. 2025
Machine Learning engineer Intern, AI Infrastructure	Los Angeles
– Built a unified vendor-recommendation platform with React frontend and Flask API gateway orchestrating Amazon Bedrock Agents, Bedrock Flows, and AWS Lambda for intelligent service provider discovery.	
– Built dual-path search architecture—internal RAG via Bedrock Knowledge Bases for certified partners and external discovery via SerpAPI (Google Maps/Yelp)—supporting parallel execution, automatic failover, and 504 timeout retries.	
– Automated infrastructure deployment including Bedrock Flow provisioning, IAM role configuration, and cross-account Lambda permissions, enabling secure execution of 1, search; 2, enrichment; and 3, ranking pipeline.	
– Enhanced search accuracy with LLM-based parameter extraction (Claude Sonnet 4) that converts natural language requests into structured Lambda inputs, complemented by trace logging to debug agent action-group invocations.	
•Research Assistant: AI-Lyric generation, code generation, music generation, and HCI	Sept. 2023 - Present
UCLA PlusLab, Los Angeles, US / C4DM, Queen Mary University of London, remote	Los Angeles
– Led the design of REFFLY, the first melody-constrained lyric revision model based on Llama 2, improving fluency and musicality by 25% over strong baselines. First-author paper accepted at NAACL 2025 (oral presentation).	
– Co-developed VDebugger, a critic-refiner system for visual program debugging. Improved error localization and correction accuracy by 3.2% across six visual reasoning benchmarks. Published at EMNLP Findings 2024.	
– Built an AI-assisted lyric writing interface based on REFFLY, with FastAPI (backend) and Vue.js + Vite (frontend); paper under submission to UIST 2025, a top-tier HCI venue.	
– Built a Variational Autoencoder (VAE) using PyTorch for symbolic music generation with interpretable latent control by regularizing latent vectors, published in Machine Intelligence Research (Impact Factor: 6.4).	
•Baidu: Spring Boot based E-commerce Service Mobile APP	June 2021 - Sept. 2021
Software Developer Intern at Baidu International Department (Do-global)	Beijing, China
– Implemented Rest API via Spring MVC including user registration/login, product listing and search, cart management, order placement, payment processing, and user profile updates.	
– Utilized Spring Data JDBC for PostgreSQL database integration, managing products, users, carts, and orders.	
– Implemented Spring Security for session-based authentication and authorization.	
– Used the Spring framework core technologies to loosely decouple all the components in the application.	
– Developed the client side using ReactJS and Ant Design, enabling users to seamlessly add items to their shopping carts and place orders.	
PROJECTS	
•MiniSpotify	
– Designed a Kotlin Spotify favorite app by fully leveraging Android Jetpack Library and Hilt Dependency Injection.	
– Implemented the BottomBar and app navigation using the Jetpack Navigation component.	
– Created a mock RESTful API with json-server and handled requests using Retrofit.	
– Built the feed/album/favorite UI in Jetpack Compose following MVVM architecture.	
– Enabled local cache for favorite features with Room Database, and integrated Google ExoPlayer for global music playback.	
•SocialAI: an AI-based social network	
– Designed and implemented a social network web application using React JS.	
– Integrated OpenAI's DALL-E 3 to assist users in creating and updating posts.	
– Improved authentication with token-based registration, login, and logout via React Router v4 and server-side JWT.	
– Developed and deployed a scalable web service in Go for post management on Google Cloud (Google App Engine).	
– Deployed ElasticSearch on GCE to enable search for recent and personal posts.	