Mengyu Zhang

Seattle, Washington, United States
6264384723 | sam.zhang.069@gmail.com

SUMMARY

Results-driven developer with experience and a diverse background in software development, machine learning, designing and implementing **high-performance, scalable systems** across industries. Proficient in **Java**, **Python**, **AWS**, and **microservices architecture** in Agile environments. Demonstrated success and experienced in building end-to-end systems, developing distributed systems, optimizing LLM models, and deploying AI solutions for real-world impact. Committed to leveraging **data-driven strategies** and **collaborative problem-solving** to deliver scalable, efficient, and innovative solutions.

CORE SKILLS

Programming: Java, JavaScript, TypeScript, Python (NumPy, Pandas, PyTorch, TensorFlow), C++ (CUDA)
Backend & Cloud: AWS (EC2, ECS, Lambda, EKS, CloudFormation, RDS, DynamoDB), Spring Boot, Flask, RESTful APIs, microservices, Terraform, Docker, Kubernetes, serverless architecture
Performance: Multithreading, multiprocessing, caching (Redis, Memcached), distributed systems, high-throughput APIs
Data: PostgreSQL, MySQL, Redshift, MongoDB, DynamoDB, Apache Kafka, AWS Kinesis, ETL automation
Other: Jenkins, CircleCI, Maven, Gradle, Git, infrastructure as code (IaC), observability (Prometheus, Grafana, AWS CloudWatch)

PROFESSIONAL EXPERIENCE

Software Engineer | *Amazon*

Seattle, WA| Jan. 2024 - Present

- Fine-tuned LLM prompts to increase API invocation accuracy from 85% to 95% for Alexa Music Expert, reducing response latency by approximately 500ms and improving model token efficiency (from 550 to 150 tokens).
- Spearheaded the development of **LLM-driven** recommendation and search features for **Alexa Radio**, enabling dynamic audio discovery and personalized station suggestions based on real-time user interaction and content metadata.
- Migrated legacy batch communication system for price change announcements with leveraging **AWS CDK** to streamline deployment, cutting infrastructure costs by 95%, reducing future operational execution time by 90%.
- Designed and developed interactive UI control page using **React** and **AWS Amplify**, enhanced system observability using **AWS X-Ray** and **CloudWatch**, providing a guided interface to manage batch communications.
- Redesigned traffic control with TPS throttling, idempotency mechanisms, and real-time traffic management using **AWS Step Functions**, increasing throughput from 150% above the baseline TPS while stabilizing downstream services.
- Automated third-part radio station names updated workflow for third Alexa Radio, enabled comprehensive harness testing for the newly added stations, decreasing manual effort by 20% and improving operational efficiency.
- Coached interns in code quality, design patterns, performance optimization, and presentation skills, ultimately boosting team productivity and technical acumen.

Software Engineer | *Open Innovation LLC (Self-Employed)*

Seattle, WA| May. 2023 – Jan. 2024

- Developed distributed web application leveraging **Node.js** and **Java microservices architecture**, ensuring scalability and modularity.
- Designed a seamless login and authentication system with **React** and **JWT**, integrating **OAuth** to provide secure Google and email-based authentication, facilitating smooth user session handling across the app.
- Integrated Stable Diffusion into application, enabling real-time text-to-image generation with **AWS Lambda** for serverless execution and **EC2** to handle concurrent high-load requests, significantly enhancing performance scalability.
- Utilized **Redis** as distributed caching solution and **Elasticsearch** to optimize data retrieval times for images and other user inputs, reducing latency and improving user experience.
- Orchestrated containerized microservices using **Docker** and deployed them via **Kubernetes**, enabling automatic horizontal scaling, fault isolation, and streamlined deployment across environments.
- Established CI/CD pipelines with **CircleCI**, automated testing using **JUnit5**, **Mockito**, and **Jest** for both unit and integration tests, ensuring continuous integration and maintaining code quality throughout the development lifecycle.

Data Science Intern | *Sikka Software*

San Jose, CA| May. 2022 – Aug. 2022

- Developed ETL Pipelines with **PySpark** and **Boto3** to update and maintenance over 1 billion patients' data stored on **AWS S3**, ensuring efficient data ingestion and preparation for downstream analytics.
- Integrated **AWS Athena** with **PyAthena** and **SQLAlchemy** to perform data management and preview on collected data.
- Built **BiLSTM** model with **PyTorch** on **AWS EC2** to predict health indicators for evaluating patients' dental health.

• Analyzed medical notes with **AWS Comprehend Medical**, performing **exploratory data analysis (EDA)** to extract key medical entities, uncovering insights into sequential procedure patterns.
• Demonstrated the feasibility of **progressive disease prediction** through time-series modeling, enabling improved patient treatment plans and early intervention strategies.

Software Engineer - AI/ML| *Inspur Group* Beijing, China| May. 2020 – Jul. 2021

• Developed **Python-based web crawler** using **Selenium WebDriver** to collect and update over **1 million images**, ensuring a robust dataset for machine learning training and evaluation.
• Designed **image preprocessing** and **augmentation filters** with **OpenCV** to adjust brightness, contrast, and apply transformations, improving model robustness and reducing overfitting during training.
• Built and deployed object detection model using **YOLOv4** with **CUDA acceleration**, implemented on **HUAWEI Bastion Host Cloud Computing Server**, achieving **95%+ accuracy** in receipt image recognition under production workloads.
• Created an **OCR invoice recognition system** with **PyTesseract**, achieving 87.4% on F1-score, **20% improvement** over baseline, by integrating preprocessing techniques for noise reduction and text segmentation.
• Designed and implemented data deduplication algorithm for makerspace-collected dataset using **fuzzy matching**, clustering techniques (**DBSCAN**), and **hash-based comparison**, reducing redundancy by **20%** and enhancing data quality for downstream analytics.
• Collaborated with cross-functional team to integrate AI solutions into production workflows, automating invoice processing and reducing manual efforts by **50%**.

PROJECT EXPERIENCE

Automated Short-Term Options Trading Bot Sep. 2022 – Jan. 2024

• Designed and implemented a Python-based trading bot for **1-day SPY/QQQ options**, achieving average monthly ROI of **20%+** using personal funds in **live market conditions** with consistent profitability.
• Integrated **Robinhood API** for real-time market data retrieval and automated order execution, implementing **fail-safes**, **rate-limiting**, and **retry logic mechanisms** to handle dynamic market volatility efficiently.
• Deployed the bot on **AWS EC2** instances with **Docker**, enabling horizontal scaling to handle concurrent trades, live analytics, and complex data processing seamlessly.
• Built a **React-based dashboard** to visualize PnL, trade performance, and live positions, integrating automated Slack/email alerts for unusual price movements or system errors efficiently.

Campus Second-Hand Platform for Students Mar. 2023 – Jul. 2023

• Developed distributed e-commerce system based on **Spring Boot** and **Spring Cloud** backend Microservice architecture.
• Created login functionality with **React** using **Firebase**, allowed **Google Authentication** and email sign-in methods.
• Integrated the application with **Redis** as distributed cache for optimizing system performance, **Elasticsearch** for faster product retrieval time, **Spring Session** for session data sharing**, thread pool** and asynchronous task for stability.
• Implemented registration/configuration center with **Nacos**, **Gateway** as gateway, and **Feign** for remote calls.
• Initialized **CI/CD Pipelines** on **Circles** and designed unit test/integration test processes with **Junit5**, **Mockito**, **Jest**.

Scholarly Text Institution Name Disambiguation |JHU - IEEE Capstone Dec. 2021 – Jun. 2022

• Implemented Python script with **Psycopg** to update database for IEEE's data lake of 5.4 million papers in **AWS Redshift**.
• Performed Named Entity Recognition with **spaCy**, created geolocator with **GeoPy** to impute geographical features.
• Applied **LightGBM** Classifier with Python to perform supervised pairwise comparison and classification.
• Deployed **Hierarchical Agglomerative Clustering** (HAC) to center and standardize clusters.
• Published algorithms and data with Sphinx documentation on AWS EC2, achieved 81.4% accuracy.

EDUCATION

Johns Hopkins University, Master of Science in Engineering, Data Science May. 2023
University of California, Los Angeles, Bachelor of Science, Applied Mathematics Major, Statistics Minor Dec. 2020
UCLA Extension, Certificate Program, Data Science Mar. 2021