Jerry Zeng

Pittsburgh, PA | 510-984-8990 | github.com/Zzz212zzZ
jerryzeng@cmu.edu | jerry.technology | linkedin.com/in/jerry-zeng-cs

Experience

Amazon, *Software Development Engineer Intern* May 2025 - Aug 2025
Amazon Stores, Amazon Fulfillment & Robotics, Sortation Team Seattle, WA

- Developed and deployed a **sorter slot configuration optimization system** across North America, delivering **12,000+ lines of production Java code** and leveraged **CloudWatch** to validate a **4.5% UPH efficiency improvement**.
- Automated daily automation updates across North American warehouses using **AWS EventBridge** to trigger **Lambda** workflows, streamlining optimization in a single run to eliminate manual updates, saving **60+ labor hours per day**.
- Built a data pipeline using **S3** and **DynamoDB** for multi-source ingestion and **AWS Glue** for transformation, enabling data-driven sorter optimization that delivered more robust allocation decisions and efficient operations.
- Implemented dynamic, region- and environment-aware deployments across NA, EU, and JP using **Google Guice** for modular injection and **Spring** for framework-level integration, ensuring scalable and reliable global rollouts.

Briefy.ai (acquired by Xmind), *Fullstack Engineer Intern* May 2024 – Aug 2024

- Built product with **TypeScript/React** (monorepo, reusable components, typed API SDK) and **Python** backend services for LLM orchestration, supporting launch as **#1 Productivity Product on Product Hunt**.
- Developed LLM pipelines to transform digital content (e.g., wiki articles, videos) into interactive mind maps and timelines in one click, forming the flagship feature later integrated into Xmind's platform.
- Deployed services on **GCP** with CI/CD (GitHub Actions, rolling releases), integrating health checks and real-time sync to ensure stable, interruption-free updates in production.

Citibank, *Fullstack Engineer Intern*, Full-time Return Offer in 2024 NG Jun 2023 – Aug 2023

- Built core microservices for a *real-time institutional bidding platform* with **Spring Boot + Kubernetes**, enabling high-throughput bid validation and pricing at **10M+ daily volume** with tenant-level rate-limiting.
- Built responsive trading dashboards with **React (TypeScript)**, and WebSockets for real-time bid updates; optimized pipeline to maintain **sub-200 ms** updates at scale and added fault-tolerant recovery for high availability.
- Built an anomaly detection pipeline leveraging **Redis** for real-time trade monitoring, integrating PyTorch models to flag irregular patterns in **sub-150 ms**; reduced false positives by 25% and strengthened trader trust.

Education

Carnegie Mellon University	M.S. in Computer Science (GPA 4.0/4.0)	Sep 2024 – May 2026
University of California, Berkeley	EECS Exchange Program (GPA 3.9/4.0)	Aug 2023 – May 2024
Chongqing University	B.S. in Software Engineering (Rank: 1/185; GPA: 93/100)	Sep 2020 – Jul 2024

Skills

Languages: JS/TS (React, Angular, Node), Python (Django, PyTorch), Java (Spring Boot), C++, Go, C#
ML/Big Data: MLflow, Pinecone, Spark, Hadoop, Kafka, Redis, PostgreSQL, MySQL, MongoDB, Neo4j
DevOps/Cloud: AWS (EC2, Lambda, DynamoDB, S3, EventBridge), GCP, Docker, Kubernetes, Nginx, Git, Terraform

Projects

Personal Full-Stack Portfolio

- Engineered a **Next.js portfolio** on **Vercel**, leveraging edge functions, serverless scaling, and a global CDN for performance; included a Warp-inspired terminal with intelligent agent/command switching.
- Designed a distributed **rate-limiting and session management system** using **Redis** (fingerprinting, token bucket, TTL counters) for high-throughput control and secure access.
- Integrated a **finetuned GPT-4.1 model** trained on a curated JSONL knowledge base to deliver contextual, domain-specific Q&A.

Santorini.games CMU 17-514 (Top Distinction Project)

- Built a 3D online board game using **React**, **Three.js**, **Spring Boot**, and **Tailwind**; applied SOLID and GoF patterns (*Factory*, *Strategy*) to develop extensible domain models and highly testable services.
- Deployed full-stack on **AWS (EC2)** behind **Nginx** with CI/CD; configured DNS with HTTPS configs; optimized asset delivery (code splitting, compression) to reduce first-load latency.

LiveStreamr — Campus-wide Live Lectures CMU 17-637 (Top Distinction Project)

- Multi-classroom one-to-many streaming via **WebRTC** (SDP offer/answer, ICE) with React; **Google OAuth** access control.
- **Django REST** + **Channels** (Redis) for **WebSocket** signaling and session control; synced live captions and on-screen bullet comments.

Publications **(Machine Learning with Application)**

Zeng, J., Xiao, F.* A High Order Fractal-based Kullback-Leibler Divergence with Application in Classification. *Expert Systems With Applications* (JCR Q1), 2024, 238:F, 122297.

Wu, Y., Liu, R., **Zeng, J.**, et al. Multi-Time Scale Aware Host Task Preferred Learning for WEEE Return Prediction. *Expert Systems With Applications* (JCR Q1), 2024, 238:E, 122160.

Liu, R., Wu, Y., **Zeng, J.**, et al. Dual Transfer Driven Multi-Source Domain Adaptation for WEEE Reverse Logistics Return Prediction. *EAI CollaborateCom*, 2023.

Zeng, J., Xiao, F.* A Fractal Belief KL Divergence for Decision Fusion. *Engineering Applications of Artificial Intelligence* (JCR Q1), 2023, 121:106027.