+

Gabriela Zeng

206-678-9506 | gabrielazengg@gmail.com | linkedin.com/in/qilin-zeng

Education

Northeastern University Seattle, WA
Master in Computer Science *Sep. 2023 – May. 2026*

Experience

SmartLink Technology May. 2025 – Sep. 2025
Software Engineer intern *China*

- Developed **scalable microservices** using **Spring Boot** and **Spring Cloud** in an event-driven architecture to support coupon issuance, redemption, and user management modules, handling over 5K transactions per minute with 99.9% uptime.
- Built a real-time event processing pipeline leveraging **Kafka Streams** and **Apache Avro** to track user activities and coupon lifecycle events across merchant systems, achieving ¡200ms message latency and 99.9% delivery accuracy.
- Designed and implemented high-throughput **RESTful APIs** with **Spring Cache** for HBase-backed CRUD operations, ensuring dynamic coupon distribution and secure user token uploads, while integrating **Hystrix** circuit breakers for fault tolerance.
- Integrated **Redis Cluster** and **Sentinel** for distributed caching of coupon token data, improving redemption performance and reducing backend database load by 20%.
- Optimized **HBase** with **Phoenix** and block caching to enhance read/write throughput and reduce disk I/O, coordinating distributed services with **ZooKeeper** for high system availability.

Around Entertainment Jun. 2024 – Sep. 2024
Software Engineer intern *Jersey City, NJ*

- Engineered a humanized AI agent capable of natural language interaction in both text and voice using **React** + **TypeScript** for the web client and **LangChain**, **FastAPI**, and **WebSockets**, powered by **OpenAI GPT-4 API** for reasoning and **Coqui TTS** for expressive speech synthesis, enabling realistic voice-based job search conversations.
- Implemented persistent conversational memory via **Redis** and **LangChain ConversationBufferMemory**, overcoming LLM statelessness to sustain multi-turn context and improving dialogue coherence by **45%** in simulated user sessions.
- Developed a **Retrieval-Augmented Generation (RAG)** pipeline leveraging **Qdrant** for vector retrieval, embedding user resumes and job descriptions with **text-embedding-3-large**, and applying fine-grained text chunking for context segmentation; achieved 35% reduction in hallucination rate and higher retrieval precision on long-context queries.
- Integrated **SerpApi** for live job-market enrichment, merging external search data with internal knowledge via hybrid ranking and semantic filtering to enhance result relevance and contextual grounding for career-matching recommendations.

Projects

Cloud-Native GoPaaS Platform Development Jan. 2025 – Apr. 2025

- Designed and built a cloud-native GoPaaS platform using **Go-micro** v3, **Kubernetes**, and **service mesh**, achieving ¡200ms inter-service latency and 99.95% uptime with features like config management, service discovery, circuit breakers, and rate limiting.
- Integrated **Istio** for secure and observable service communication with automatic mTLS, traffic shifting, and retries; accelerated release cycles by 3× via **Helm**-based deployment and rollback automation.
- Architected microservices using **gRPC** and **protobuf** for efficient inter-service communication; implemented **API Gateway** with **Hystrix** circuit breakers, load balancing, and rate limiting, reducing system failure rates by 35% under peak traffic.
- Managed scalable **Kubernetes** clusters with **Dockerized** services, **NGINX** Ingress, and custom controllers for traffic routing and job scheduling; enhanced observability and tuning using **Prometheus** and **Grafana**, supporting 30% traffic growth.

Distributed LiftRide Analytics Platform Sep. 2024 – Dec. 2024

- Engineered a distributed, event-driven analytics system to capture high-volume skier ride data using **Node.js**, **RabbitMQ**, and **AWS DynamoDB**, achieving horizontal scalability through **partitioned** writes, sharded queues, and clustered consumers across multiple availability zones.
- Implemented asynchronous microservices with **AWS Elastic Load Balancer**, **Redis Cluster**, and cache-aside patterns to deliver sub-200 ms query latency and fault-tolerant request handling under peak load.
- Optimized data modeling with **DynamoDB GSIs** and time-bucketed sort keys for parallel reads/writes, ensuring consistency and durability through **replication** and worker coordination. Stored aggregated ride summaries in **MongoDB** for internal reporting and performance analysis.

Technical Skills

Languages: Java, Python, TypeScript, C, C++, JavaScript, SQL, MATLAB, HTML, CSS
Frameworks: React, FastAPI, LangChain, Node.js, Spring Boot, Go-micro, gRPC, Kafka Streams
Databases & Messaging: MongoDB, DynamoDB, Redis, HBase, Qdrant, MySQL, PostgreSQL, RabbitMQ
Systems & Tools: Docker, Kubernetes, AWS, Helm, Prometheus, Git, Linux