

Molly Zeng

510-714-9297 | yumengzeng.molly@gmail.com | linkedin.com/in/mollyzeng | github.com/lemonzeng

EXPERIENCE

Alibaba International — AI Design Platform

Hangzhou, China

Backend Engineering Intern

May 2025 – Aug. 2025

- Designed an asynchronous execution pipeline handling **100k+ daily requests**, maintaining **<100ms** API latency by validating submissions, enqueueing GPU-bound tasks to RocketMQ, and decoupling request handling from execution.
- Engineered a **distributed result routing service** consuming RocketMQ broadcasts to map execution results to the correct WebSocket sessions across backend instances, enabling real-time progress updates for workflow runs.
- Reduced workflow integration effort by **50%** through a shared training pipeline architecture, eliminating duplicated GPU initialization and simplifying onboarding for new processing workflows.
- Introduced per-epoch checkpoint polling to enable mid-training output preview and earlier failure detection for long-running asynchronous jobs, reducing wasted GPU runtime.

Seek AI — Discover AI Tools

San Francisco, CA

Backend Software Engineer Intern

Dec. 2023 – Mar. 2024

- Led development of a fault-tolerant data ingestion pipeline aggregating 3,700+ AI tools across 50+ domains using concurrent workers, proxy rotation, and checkpointing, achieving **<0.5% failure rate** in production scraping workloads and supporting a platform ranked **#3 Product of the Day on Product Hunt**.
- Reduced database I/O by **75%** by redesigning analytics storage from per-event rows to aggregated session documents, enabling O(1) session retrieval across 50,000+ search logs to power tool ranking algorithms.
- Developed a Flask microservice managing the Elasticsearch index with exponential backoff retry (3 attempts, 10s cap), ensuring newly submitted tools are indexed and discoverable in semantic search without manual intervention.
- Enforced cross-collection write consistency for user interactions using MongoDB multi-document ACID transactions, guaranteeing atomic like-count updates and saved-list state under concurrent writes with automatic rollback.

PROJECTS

Distributed Live Chat System | Go, AWS, DynamoDB, Redis, SNS/SQS, Terraform

Jan. 2026 – Mar. 2026

- Cut reaction POST p95 latency by **74% (310ms → 80ms)** and increased throughput by **44%** (387 → 558 req/s) by replacing synchronous DynamoDB writes with an SQS-backed async aggregation pipeline.
- Re-architected message delivery from polling to persistent WebSocket connections, reducing p50 latency **5.4× (970ms → 180ms)** and improving real-time responsiveness.
- Designed cross-instance message fan-out using SNS→SQS broadcast and Redis-based session caching across replicated Go services behind an ALB, ensuring reliable message delivery in horizontally scaled deployments.

Scalable LLM Inference Optimization | vLLM, AWS SageMaker, Vertex AI, Docker

Mar. 2026 – Apr. 2026

- Designed and evaluated LLM inference serving pipelines across local and cloud GPU environments (RTX 2080, AWS SageMaker A10G, Vertex AI L4), identifying hardware-aware deployment tradeoffs that delivered up to **2.1× lower latency** on A10G versus FP16 baselines.
- Improved inference serving performance by replacing a HuggingFace + FastAPI baseline with vLLM, reducing average latency by **32.7%**, lowering P95 latency by **37.1%**, and scaling throughput from **59 → 512 tok/s**.

EDUCATION

Northeastern University

Oakland, CA

M.S. in Computer Science, GPA: 4.0/4.0

Sep. 2023 – May 2026

Zhejiang University

Hangzhou, China

B.A. in Accounting (Honors Program), GPA: 3.65/4.0

Sep. 2018 – Jul. 2022

TECHNICAL SKILLS

Languages: Python, Java, Go, SQL, TypeScript, C++

Frameworks: FastAPI, Spring Boot, Gin, Flask

Data & Messaging: PostgreSQL, MongoDB, DynamoDB, Redis, Kafka, RocketMQ, Elasticsearch

Infrastructure: AWS (ECS, SNS, SQS, S3, SageMaker), GCP (Vertex AI), Docker, Terraform, Linux, CI/CD