Guanyu Zhang

evanz1627@gmail.com guanyu-zhang.github.io guanyuzhang guanyu-zhang

EDUCATION

Columbia University New York, NY
M.S. in Computer Science, Machine Learning Track (Course Assistant: COMS 6156 Cloud Computing) Dec 2022
The University of Texas at Austin Austin, TX
B.S. in Mathematics (Honors: College Scholars) Aug 2021

WORK EXPERIENCE

Ant International (Ant Group) Sunnyvale, CA
Software Engineer Jan 2024 – present

- Developed and deployed an AI-powered testing framework combining **RAG** and **LoRA**-tuned Qwen3-30B-A3B: generated **SFT** data distilled from multiple large language models (with CoT) on code–test pairs, and applied markdown heading–based **chunking** in the **RAG** retrieval layer to automate end-to-end test case generation.
- Built a production-grade in-memory cache framework in **Java**, decoupling legacy DAO dependencies for reusability and adding **transaction** support for atomic cache operations, supporting **over $10B GMV** and **2M daily transactions**.
- Developed a native web-search tool in **Golang** for a multi-agent system, deploying **metasearch engines** and **crawler servers** with **Kubernetes** to improve latency performance and reduce tool API failure rate.
- Optimized Amazon compliance APIs using **Kafka** to handle asynchronous database writes and **Airflow** and **Cassandra** for data aggregation and storage, reducing response latency by **50%** and increasing API availability under peak load.
- Executed tenant-level database migration from **Alicloud** to **Google Cloud Platform** using **Jenkins**, **Flink**, and **Kafka**, automating data synchronization to ensure consistency and minimize downtime during cutover.

Alipay US, Inc. (Ant Group) Sunnyvale, CA
Software Engineer Jan 2023 – Dec 2023

- Implemented a cross-border bank account service using **Java Spring Boot** with **distributed transaction processing**, supporting over **$12B annual account transfer volume** across more than **15 regions** with zero transaction fee.
- Engineered a payment orchestration layer for Venmo and PayPal integration on AliExpress, scaling a **Kubernetes**-based 500-pod microservice system to handle **3,000–5,000 QPS** and serve millions of daily buyers with consistent reliability.
- Rebuilt Alipay's Shopify payment integration by replacing legacy SDKs with new API-based payment interfaces, launching wallet and card payment apps with **Redis** caching and **asynchronous message queues**, meeting Shopify's integration requirements and serving over **50,000** global merchants with high availability.
- Implemented a **CI/CD** pipeline with **Jenkins** for the digital wallet frontend using **JavaScript**, automating build and deployment workflows to improve release efficiency and reduce manual interventions.

Qingdao Anzhi Capital Management Shanghai, China
Software Engineer Internship Oct 2020 – Jan 2021

- Built and maintained a cloud-based trading data infrastructure on **AWS**, designing distributed pipelines for transaction data ingestion, storage, and retrieval across multiple markets. Deployed **Amazon RDS** and **DynamoDB** with automated backup, replication, and failover policies to ensure high availability and low-latency access for daily trading operations.
- Developed a **machine learning–driven** alpha-factor generation system in **Python**, focusing on **feature engineering** using multi-resolution **OHLC** (Open-High-Low-Close) data from equities held by mutual funds. Engineered features from minute-, daily-, and monthly-level **K-line data** to capture both short-term momentum and long-term trend signals.
- Implemented and evaluated two predictive models — a **ResNet-based deep model** and an **ensemble of linear models** — using **PyTorch** and **scikit-learn** to forecast **short-term index-option movement** (typically hourly intervals). Visualized model outputs and exposed results through an internal API consumed by the quantitative-trading team as part of their alpha-factor library.

PROJECTS

Lobe Chat Open-Source Contribution Sep 2025 – Present

- Contributed to Lobe Chat, an open-source AI chat framework built with **TypeScript** for private deployment, by developing agent tools.

Speech Data Annotation Reliability Project Nov 2020 – Dec 2020

- Developed a pipeline using **MFCC** features and pre-trained **ResNet** models to process raw speech data, and built a **VGG-S** classifier that improved **F1 score by 5%** over the RF baseline.

TECHNICAL SKILLS

C++, Go, Java, Python, JavaScript, SQL; AWS & GCP; MySQL, MongoDB, Redis; Docker, Kubernetes; PyTorch, scikit-learn, Transformers, PySpark