Frank Li

angli23@cs.washington.edu | linkedin.com/in/anglifrank | github.com/frangkli | frangkli.com

Education

University of Washington – Seattle, B.S. Computer Science (GPA: 3.9) Graduation: Spring 2026

Courses: Data Structures & Parallelism, Algorithms, Distributed Systems, Datacenter Systems, CV, NLP, ML, AI, Deep Learning, Compilers, Databases, Operating Systems, Linear Algebra, High-Performance Computing, Linear Optimization, ML Systems, Quantum Computation

Technical Skills

Languages:	Python, C, C++, Java, Kotlin, Rust, HTML, CSS, Javascript, Typescript, Shell, SQL, Go, SystemVerilog, OCaml, Assembly
Frameworks:	Protobuf, gRPC, MPI, CUDA, NumPy, Jax, Pandas, Arrow, PyTorch, Ray, MCP, AWS CDK, Node/Bun, React, LangChain
Tools:	Linux, Git, CMake, PostgreSQL, PGVector, AWS, Kafka, Redis, Docker, Kubernetes, GDB, Parquet/Avro, Iceberg, DuckDB

Experiences

NVIDIA Fall 2025
Incoming Systems Software Engineer Intern - MLOps for Autonomous Vehicles Santa Clara, CA

Two Sigma Jun 2025 – Aug 2025
Software Engineer Intern - Systematic Macro Engineering New York, NY

- Developed a Python **ETL framework** and UI for analyzing **large time-series trading logs** using **Arrow and Iceberg**, reusable firm-wide.
- **Enabled half-day trading** for our 0DTE options pipeline by iterating with QRs to identify model assumptions and analyze quant diff.
- Created a **record-replay system** to capture and visualize simulation system metrics in Grafana, allowing devs to **identify bottlenecks**.
- Enhanced the internal coding agent by extending it with a **RAG system for sharing conversation memory**, accessible via agent MCP.

Amazon Jun 2023 – Jun 2025
Jr. Software Development Engineer III (Year-round SDE intern) Seattle, WA

- Developed **dozens of new features** for our tier-1 content management service with **millions of enterprise users** using **Java and AWS**.
- Improved user privacy and experience by **automating end-to-end data encryption and asset regionalization, reducing latency by 20%**.
- Architected a **generic end-to-end testing system** and a serverless backend, both **reusable via infrastructure-as-code** with AWS CDK.
- Led the restructuring of data models and **utilized concurrent asynchronous queries** to **reduce tail latency by 40%** for our customers.

UW High-Performance and Data-Intensive Computing (HPDIC) Lab Jan 2024 – Jun 2024
Undergraduate Researcher (Vector Databases) Seattle, WA

- Formulated **two new multi-vector search query algorithms** based on hierarchical navigable small-world and custom distance metrics.
- Extended pgvector (PostgreSQL vector extension with **9.8k Github stars written in C**) to support semantic multi-vector queries.
- Developed and benchmarked a **Python ORM** with new operators on **HPC clusters**, maintaining similar latency to single-vector queries.

Shanghai Media Intelligence Technology Jul 2021 – Aug 2021
Software Engineer Intern Shanghai, China

- Developed a presentation tool with React that compares high-definition video streams with **timestamp synchronization and caching**.
- Enhanced the data pipeline by **implementing additional automation and data augmentation, increasing model accuracy by 5%**.

Creative Hose Equipment Technology Jul 2018 – Aug 2018
Fullstack Developer Intern Beijing, China

- Developed tools for supply line inventory management using **Java with Spring Boot, Hibernate, and JSP with Oracle database**.
- **Secured data operations** by utilizing prepared statements and serializable transactions to prevent SQL injection and race conditions.
- Spearheaded the transition to **embedded PWA** by creating a prototype, **improving performance and increasing adoptability by 40%**.

Projects

GitHub Analysis MCP Tool | *Python, uv, Model Context Protocol, Ollama, pytest, Qwen-2.5*

- Leveraged the **Model Content Protocol** to build a MCP server that allows AI agents to query repositories and analyze their content.
- Developed a CLI client with Ollama to allow Qwen-2.5 to autonomously choose and use MCP tools based on user prompts.

Prefill-Decode Disaggregation System for LLM Inference | *Python, Ray Data, PyTorch, Transformers, asyncio*

- Engineered an actor-based LLM serving system with **Ray** with KV cache and **disaggregated prefill and decode inference phases**.

Cloud VM CPU Cache Latency Analyzer | *C++, Python, GCP, Shell, Pandas, Seaborn*

- Leveraged **C++ multithreading, atomics, and syscalls** to set threads' CPU core affinities, analyzing **latency of cache coherent operations**.
- Presented the read and write latencies between each cores on a heatmap, identifying CPU core pairings that **reduce latencies by 80%**.

Dockerized Yelp Clone with Performance Analysis | *Go, Lua, gRPC, Protobuf, Docker, Kubernetes, Shell, GCP*

- Architected a Yelp clone with dockerized microservices in **Go**, utilizing **gRPC** for communication and **Kubernetes** for orchestration.
- Coded **Lua scripts with wrk2** to benchmark the system and **identify bottlenecks**, then implemented caching to **reduce latency by 30%**.