Charlton Shih

LinkedIn | GitHub | charltonshih645@g.ucla.edu | (951) 905-4920

Education

Applying to MS/PhD programs **September 2027**

University of California, Los Angeles (UCLA) **September 2023 - June 2027**
B.S. Computer Science *Los Angeles, CA*

- **Relevant Coursework:** Operating Systems, Networks, Databases, Computer Architecture, Data Structures & Algorithms, Machine Learning, Natural Language Processing, Software Engineering, Deep Learning for Computer Vision

Technical Skills

Languages: Python, Go, Rust, C/C++, Java, SQL, TypeScript, JavaScript, HTML/CSS
Frameworks & Libraries: FastAPI, Django, gRPC, Node.js, LangGraph, OpenCV, React
Machine Learning & Data Science: PyTorch, TensorFlow, Scikit-learn, XGBoost, NumPy
Databases: PostgreSQL, Redis, Pinecone, MongoDB, Google Firebase
Cloud & Infrastructure: AWS, Google Cloud, Linux/Unix, Docker, Kubernetes, Terraform, RabbitMQ, Git

Experience

NBCUniversal September 2026 – Present
Software Engineer Intern *University City, CA*

- Reduce platform health review time by **25%** by developing an admin-only **observability dashboard** with **role-based access control** and a service topology map that helps administrators locate unhealthy services.
- Centralize health monitoring for **10+** services by building a read-only **API** over **DynamoDB** projection tables, serving aggregated summaries while retaining raw telemetry in **Datadog** and **CloudWatch**.
- Deliver fixes and enhancements across the Legal AI platform by resolving **Jira** tickets involving **RAG**, **LangChain**, **Docker**, and **AWS Lambda, API Gateway, CDK, and CloudFront**.

AfterQuery October 2025 – April 2026
Software Engineer Intern *San Francisco, CA*

- Developed LLM evaluation infrastructure on **AWS** with **Python**, **Redis**, and **Terraform**, running **750+** concurrent evaluation jobs and **150+** parallel batches through autoscaling and **CI/CD** pipelines to benchmark model accuracy.
- Built a **FastAPI** backend with **40+** endpoints for long-running agent evaluations, saving checkpoints in **PostgreSQL** to recover interrupted runs and streaming progress events through **RabbitMQ** to a **Redux** dashboard for real-time execution monitoring.
- Implemented a Model Context Protocol (**MCP**) server exposing **15+** tool-call APIs through containerized services, giving AI agent runtimes access to backend resources and automating end-to-end evaluations across **100+** agent tasks.

Bruinwalk (UCLA Student Media) October 2025 – Present
Software Engineer Intern *Los Angeles, CA*

- Implemented a **Redis** caching layer for enrollment and faculty data ingestion, reducing redundant API traffic by **90%**, lowering network bandwidth usage, and improving ingestion throughput for high-frequency dataset updates.
- Refactored core APIs and **Django** backend services supporting **1M+** monthly site visits, automating data sanitization and annotation workflows through **4+** production admin dashboards and improving system reliability by **25%**.
- Optimized **Kubernetes** container orchestration and the **Go gRPC** communication layer, reducing request latency by **60ms** under production load and improving service responsiveness for **42,000+** users.

Research Experience

Toyota Research Institute April 2026 – Present
Student Researcher *Los Angeles, CA*

- Built an agentic **DAgger** loop using LLMs with privileged scenario data to refine driving trajectories; fine-tuned **DrivoR** on large datasets using DAgger, ensemble DAgger, ThriftyDAgger, and AIM for human-like driving.
- Created **NexusSim** with **CUDA**, **Isaac Sim**, and **Isaac Lab** for scenario simulation and 3D asset generation; integrated **gRPC**/NuRec rendering of real-world 3D Gaussian Splatting (**3DGS**) scenes and added **CI** regression test suites.

Leadership

TappedIn (Creative Labs at UCLA) March 2026 – Present
Lead AI/Backend Developer *Los Angeles, CA*

- Led a 7-person team building a retrieval-augmented generation (RAG) pipeline with LangGraph, OpenAI, and Pinecone for LA event discovery; scoped deliverables and ran biweekly sprints to ship the platform within one quarter.

Samueli Engineering School UCLA March 2025 – June 2025
Teaching Assistant - Software Construction (CS 35L) *Los Angeles, CA*

- Taught systems programming, networking, OS concepts, security, and software testing in 20–30-student discussion sections and office hours; supported software construction coursework and projects in collaboration with course professors.