Shouyi Li

sl5632@columbia.edu | +1 (323)-633-8029 | linkedin.com/in/shouyili

EDUCATION

Columbia University — New York, NY
- MS in Computer Science (Machine Learning Track) — Aug 2024 – Dec 2025

University of Southern California — Los Angeles, CA
- BS in Computer Engineering and Computer Science, GPA: 3.9/4.0 — Aug 2020 – May 2024

SKILLS

Languages: C++/C, Python, Java, HTML & CSS, Javascript/Typescript, SQL, Verilog
Frameworks/Libraries: Spring/Spring-boot, REST API, FastAPI, ReactJS, Flask, jUnit, AJAX
Cloud/Databases: AWS (EC2, RDS), GCP, MySQL, NoSQL, PostgreSQL, Google Firebase, MongoDB, Terraform
Tools: Linux, Docker, Kubernetes, Bash/Shell, VM, PyCharm, Android Studio, DataGrip, MapReduce, Git, CI/CD
Courses/Concepts: Data Structures, Algorithms, OOP, Networking, Operating System, Database, Web development, Cloud Computing, Embedded Systems, Parallel and Distributed Systems, Multi-threading, ML, Agile, Waterfall, Scrum

PROFESSIONAL EXPERIENCES

Meituan — Beijing, China
Software Engineer (AI/ML) Intern — May 2024 – Aug 2024
- Optimized LLM inference time to improve prompt response speed by 23%, significantly enhanced API performance for large language model applications such as customer service chatbots and
- Implemented batch inference optimization for large-scale LLM deployment based on research literatures, used dynamic pruning techniques with a parallel-processed predictor to reduce MLP computations by up to 50%
- Collaborated with cross-functional teams for end-to-end deployment and facilitating downstream communications

Sunwood Ecological Engineering — Irvine, CA
Software Engineering Intern — May 2023 – Aug 2023
- Engaged in an internal system development to help workers manage ecological project and track progress
- Developed a responsive React frontend, leveraging useMemo and useCallback hooks to optimize rendering and improve application performance, reducing unnecessary re-renders by 40%
- Implemented REST APIs using Spring Boot, incorporating Spring Security to handle authentication and authorization, ensuring secure data access across the platform
- Optimized PostgreSQL database queries by refining query structures, improved data retrieval speeds by 30%
- Containerized and deployed applications using Docker, managed deployment and scaling on AWS EC2 instances
- Used Terraform to automate deployment workflows, collaborated on GitHub for version control and CI/CD

Jensen-Group Technology — Xuzhou, China
Software Engineering Intern — May 2022 – Aug 2022
- Implemented REST API endpoints to allow workers to manage industrial laundry device production pipeline data
- Built portions of a NoSQL database using MongoDB for storing and analyzing historical production data
- Collaborated with cross-functional teams, including hardware engineers and data analysts, to align backend integrations with production pipeline requirements for large industrial devices

Hongmeng Measurement and Control Technology — Xuzhou, China
Mobile Software Engineering Intern — Feb 2021 – Aug 2021
- Achieved real-time sensor data transmission and monitoring for mobile device through Android APP
- Implemented the Modbus protocol with CRC-16 checksum algorithm to ensure 100% error bit detection, optimizing performance by accelerating CRC computation by an average of 5.3 times with a lookup table approach

University of Southern California, High Performance Computing Lab and CSI Lab — Los Angeles, CA
Student Researcher — Mar 2023 – Aug 2024
- Engaged in project "*Accelerating ViT Inference on FPGA through Static and Dynamic Pruning*" (FCCM 2024)
- Conducted "*Investigating the Efficiency and Performance Gains of FPGA-accelerated CNN*" (CONF-MLA 2023)
- Developed project "*Deep Morphological Profiling of Immune Cells in Peripheral Blood*"

PROJECTS

CampusHub — Aug 2024 – Dec 2024
- Designed a microservices-based school event app to allows users to create, join, and manage campus events
- Utilized React and FastAPI, containerized with Docker and led deployment on AWS EC2 instances
- Integrated MySQL RDS with AWS DBaaS, enabling replication and indexing for scalable data management

Planner Pal — Sept 2022 – Dec 2022
- Developed a multifunctional Chrome extension, helped over 200 users optimize with day-to-day tasks schedule
- Built backend APIs using Spring-boot, with Javascript as frontend, Google Firebase as database, and led deployment on Google Compute Engine; engaged with users to gather feedback and enhance app's features