Richard So

➤ Seattle, WA | ☎ 347-281-3815 | ✉ richardso2021@gmail.com | in/richardso21 | 🌐 sorichard.com

Education

Georgia Institute of Technology 08/2024 - 05/2025

M.S. Computer Science, Machine Learning - GPA: 4.0 *Atlanta, GA*

- Coursework: ML, Deep Learning, Computer Vision, NLP, Databases, Networks, Algorithms Honors

Georgia Institute of Technology 08/2021 - 05/2024

B.S. Computer Science - GPA: 4.0 *Atlanta, GA*

Work Experience

Amazon | *Software Development Engineer* 08/2025 - Present

- Maintained **data lake infrastructure** (AWS Glue, EMR, Lambda, S3) for Amazon Brand Analytics, enabling sellers to monitor **sales performance insights across 2B+ products**.
- Orchestrated **high-throughput data vending pipelines** with Apache Spark and Airflow to transform **~100 TB/week** of raw purchase activity into curated analytical datasets for internal teams and sellers.
- Developed core components of an **explainable ML root cause analysis service**, leveraging **Shapley value feature attribution** to identify drivers of product underperformance relative to benchmarks.

Data To Insights (D2I) Lab @ Georgia Tech | *Research Assistant* 01/2025 - 05/2025

- Investigated time-to-first-token (TTFT) reduction in **multi-tenant LLMs** by overlapping document retrieval and prefill.
- Led **scenario-based benchmarking** by simulating web crawler and ANNS retrieval to evaluate latency improvements of our custom token prefilling mechanism under realistic workloads.
- Analyzed results for and **co-authored MLSys research paper** demonstrating up to **11x faster latencies in TTFT**.

Amazon Web Services | *Software Engineering Intern (ML)* 05/2024 - 08/2024

- Implemented **ML-driven analysis of developer workflows** by aggregating behavioral features (keystrokes, UI events, time-on-task), enabling quantitative DX comparison across cloud platforms and **influencing AWS roadmap decisions**.
- Re-architected a batch inference pipeline to exploit parallelism, **reducing runtime by >85% (hours → minutes)**.
- Automated developer intent and task classification from desktop screenshots using AWS **Rekognition**, **Textract**, and **multi-modal LLM prompting on Bedrock**.

Tanium | *Software Engineering Intern* 06/2023 - 08/2023

- Built CRUD logging to Tanium console's PostgreSQL backend and REST endpoints to support a **customer-facing audit feature**, allowing review of console activity and detection of unauthorized configuration changes.
- Resolved **50+ feature/bug tickets** within a 10-week internship maintaining a Knex.js + React TypeScript codebase.
- Ensured code quality and correctness by applying **TDD & validation best practices** using Jest, Jasmine, and Joi.

Projects

Generative Data Augmentation for Image Classification 04/2024

- Experimented with **Stable Diffusion** and **ControlNet** to enhance image classification accuracy when data is scarce.
- Observed a **10% F1 increase** for Resnet-50 on a limited dataset when augmented with ControlNet-generated images.
- Orchestrated **large-scale DL experiments**, sweeping across 100+ hyperparameter and model configurations.

LC3Tools 09/2023 - 05/2024

- **Lead maintainer** of the educational tooling suite for coding, assembling, and simulating LC-3 assembly programs.
- Actively collected student & instructor feedback to continuously drive **major quality-of-life enhancements**.
- Served **1000+ students every semester** as one of the core, required tools for Georgia Tech's CS2110 course.

Skills

Programming Languages	Python, TypeScript/JavaScript, Java, Scala, C/C++, Go, Lua
Frameworks & Libraries	NumPy, Pandas, SkLearn, PyTorch, Lightning, Apache Spark, React, Svelte
Databases & Misc.	PostgreSQL, SQLite, DynamoDB, Elasticsearch, Docker, Apache Airflow, AWS CDK