Ari Singer

✉ me@arisinger.net | 📞 +1 248-763-8364

in LinkedIn | GitHub | 🌐 arisinger.dev

🏠 New York, NY | </> Resume Source

EXPERIENCE

AMAZON   New York, NY

Software Development Engineer II   JUNE 2025 — PRESENT

- Built a production grade LLM inference platform as primary IC alongside principal engineers (L7/L8) — disaggregated serving with Nvidia Dynamo, KV-aware routing, and a custom orchestration layer supporting A/B deployments, fault tolerance, and rate limiting, fully IaC on Kubernetes (AWS CDK)
- Owned production hardening including observability (metrics, dashboards, alarms), security compliance, air-gapped VPC networking, and open-source dependency mirroring
- Shipped an agent playground (Streamlit) and demo agents (LangChain) to showcase model capabilities to leadership
- Supported post-training and evaluation efforts — ran SFT jobs with Verl and debugged training pipeline issues for research scientists

Software Development Engineer I   MARCH 2024 — JUNE 2025

- Implemented an asynchronous pipeline that routed 100K+ ad creatives through AI-based transformation services, converting assets into new format-compliant variants — unlocked hundreds of millions of new DSP impressions and substantial incremental annual revenue

Software Development Engineer Intern (Seattle, WA)   MAY 2022 — AUGUST 2022

- Automated QA testing with a pipeline using SQS, Lambda, and AWS Device Farm to capture ad asset interactions (taps, swipes, presses) and push screen recordings to S3 — reduced QA cycle time from hours to minutes
- Created an internal web GUI for QA testers to submit automation jobs and review results, replacing a fully manual process

Software Development Engineer Intern (Remote)   MAY 2021 — AUGUST 2021

- Developed an automation layer that pulled Fire TV and Fire Tablet ad campaign data from S3, transformed and enriched the data, and wrote results to DynamoDB

UNIVERSITY OF MICHIGAN   Ann Arbor, MI

Graduate Student Instructor   AUGUST 2023 — MAY 2024

- EECS 441 — Mobile App Development: grading, lectures, office hours

Graduate Researcher   JUNE 2023 — MAY 2024

- Developed and deployed agentic LLM applications for K–12 classrooms before frameworks like LangChain existed — implemented chain-of-thought prompting and context management, deployed to real classrooms with the Center for Digital Curricula

Undergraduate Researcher   JANUARY 2021 — DECEMBER 2021

- Studied reinforcement learning techniques for tuning inkjet printer parameters within a multidisciplinary team

PROJECTS

SMART DOOR HANDLE   GitHub

- Built a smart door handle with custom PCB for automatic unlock/locking based on user proximity + capacitive touch along with a companion app for remote access. Handled all embedded programming, cloud backend, integration with the iOS app.

EDUCATION

UNIVERSITY OF MICHIGAN   Ann Arbor, MI

Master of Science in Engineering in Computer Science   May 2024

- GPA: 3.90/4.00

Bachelor of Science in Engineering in Computer Science, Mathematics Minor   May 2023

- GPA: 3.95/4.00 (Summa Cum Laude)

Relevant Coursework: Systems of Generative AI, Machine Learning, Artificial Intelligence, Natural Language Processing, Operating Systems, Compilers, Computer Architecture, Cryptography, Computer Networks, Web Systems, Embedded Systems, Logic Design

SKILLS AND TECHNOLOGIES

LANGUAGES	Python, Java, C++, C, Typescript/Javascript, Bash, SQL
INFRASTRUCTURE	AWS, Kubernetes, Docker, Linux, Git, vLLM, Nvidia Dynamo, Verl
FRAMEWORKS	LangChain, Streamlit, React.js, pytest