Matthew Kuo

github.com/kuomat | www.linkedin.com/in/kuomat/

Email : mkuo@seas.upenn.edu
Mobile : +1 (626) 517-6009

Education

University of Pennsylvania — Philadelphia, PA

Master of Engineering in Robotics; GPA: 3.9/4.0 — *May 2026*

Bachelor of Science in Engineering (Major: Computer Science; Minor: Statistics) — *May 2026*

- **Relevant Coursework:** Algorithms & Data Structures; Big Data & Data Science; Operating Systems; Internet & Web System; Relational Databases; Linear Algebra; Scalable Cloud Computing; Statistics; Discrete Mathematics; Intro to Robotics

Experience

Software Developer Intern — May 2025 – Aug. 2025

Amazon | Java, TypeScript, AWS (Bedrock, Lambda, CloudFormation) — *Bellevue, WA*

- Architected a **GenAI-powered support agent** using **Model Context Protocol (MCP)** and **Bedrock**, automating ticket triage for 250+ engineers and slashing resolution time by **80%**
- Designed **serverless ETL pipelines** processing **350M+ daily checkouts** via **Lambda** & **S3** with zero operational overhead
- Codified infrastructure using **CloudFormation** and CI/CD pipelines, reducing environment provisioning time by **90%** and eliminating **50+** manual configuration steps
- Enhanced system observability by deploying centralized **CloudWatch** dashboards, decreasing Mean Time to Detect (MTTD) incidents by **70%**

Deep Learning Researcher — Mar. 2024 – Present

Google X AI Research Lab (Prof. Mayur Naik) | PyTorch, JAX, Bash — *Philadelphia, PA*

- Co-authored NeurIPS 2025 **Spotlight paper** (top 3%) introducing **SGClip**, a foundation model generating spatio-temporal scene graphs for embodied agents
- Engineered self-supervised training pipeline in **PyTorch**, reducing visual errors by **39%** in multi-modal agents
- Re-architected inference pipeline using **JAX**, cutting latency by **25%**; work featured on the **Google Open Source Blog**
- Scaled dataset coverage by **7x** by engineering an automated labeling pipeline using **VideoLLaMA + SAM2**
- Accelerated training by **3x** via distributed training across 9 servers using **PyTorch DistributedDataParallel**

Software Developer Intern — Jun. 2024 – Aug. 2024

BizzyBots (Wharton-backed startup) | TypeScript, tRPC, GCP, Redis — *Philadelphia, PA*

- Engineered a **GPT-powered** negotiation simulator web app, scaling to 70+ active users and handling real-time dynamic pricing scenarios on **GCP**
- Refactored backend architecture by migrating 20+ endpoints to **tRPC**, enforcing **end-to-end type safety**
- Slashed database read latency by **35%** by implementing a **Redis** look-aside cache for **Firestore** conversation history

Head Teaching Assistant — Jun. 2024 – Dec. 2024

CIS 7000: Large Language Models | PyTorch, HuggingFace — *Philadelphia, PA*

- Designed advanced curriculum modules on **Reinforcement Learning (RLHF)** and **BERT** attention, translating complex research papers into interactive **PyTorch** demos
- Architected graduate-level coding projects implementing **Transformers**, **LoRA**, and **SliceGPT**, enabling 50+ PhD students to optimize LLMs for efficient inference

Projects

Bytenet Search | *Java, AWS (EC2, S3), Distributed System Design*

- Built a distributed **search engine** indexing 200k+ pages and serving search queries with latency < 1000ms on AWS EC2
- Crawled and ranked **200k+** web pages on EC2, using TF-IDF, PageRank, and combined scoring to optimize search results

Insta Lite | *JavaScript, Apache Spark, Apache Kafka, Page Rank, TCP/IP*

- Built a MERN app that serves **REST APIs** to enable users to create/view posts and manage friend requests
- Implemented scalable data storage architecture using AWS RDS for numerical data, S3 for images, ChromaDB for image embeddings, and Spark to parallelize the **pagerank algorithm**
- Established a **Kafka platform** for real-time streaming and processing of user-generated content, and built a **recommendation system** to suggest new friends based on current connections

DataGuard | *Python, Apache Kafka, Docker, PII Detection*

- Built a **real-time Kafka data firewall** with confidence-based PII detection (credit cards, SSNs, emails, phones, IPs) using **Luhn validation** and context analysis, reducing false positives by **40%+**
- Implemented **fail-closed semantics** with deterministic message IDs, ensuring zero data leakage and enabling safe replays; deployed via **Docker Compose** with comprehensive audit trails

Skills

Languages: Python, Java, TypeScript, JavaScript, SQL, C
Frameworks/Tools: Docker, Git, Jenkins, React, Node.js, Express.js, Flask, Django, PyTorch, TensorFlow, Apache Spark, Kafka
Cloud/Data Technologies: AWS (EC2, VPC, RDS, S3), GCP, PostgreSQL, MongoDB, MySQL, SQLAlchemy, Redis, DynamoDB