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

Skills

Languages: Java, Python, 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

Experience

Software Developer Intern — May. 2025 – Aug. 2025

Amazon | Java, TypeScript, Model Context Protocol — *Bellevue, WA*

- Built high-throughput **MCP servers** and a Bedrock-powered chatbot to surface actionable insights, cutting ticket triage time by 80% for 250+ engineers
- Designed a **serverless** ETL and logging pipeline for 350M+ daily checkouts using Lambda and S3, enabling scalable ingestion with **zero ops overhead**

Deep Learning Researcher — Mar. 2024 – Present

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

- Second author of a NeurIPS (2025) submission introducing **SGClip**, a **CLIP**-based model that generates structured scene graphs for visual understanding in embodied agents
- Built a self-supervised training pipeline (PyTorch), reducing visual errors by **39%** in multi-modal agents like **GPT-4o**
- Developed a **probabilistic scene graph** representation to handle uncertainty and enable fine-grained visual reasoning
- Expanded dataset coverage by **7x** via an automated **VideoLLaMA + SAM2** pipeline with sliding window tracking
- Accelerated training by **3x** through distributed multi-threading across 9 servers using **DistributedDataParallel (DDP)**

Software Developer Intern — Jun. 2024 – Aug. 2024

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

- Enabled 70+ users to simulate dynamic price negotiations by deploying a **GPT**-powered chatbot web app
- Boosted offer relevance by **30%** by designing a real-time conversation filtering algorithm, tested via **Jest** in **CI/CD**
- Migrated 20+ HTTP endpoints to tRPC, eliminating client-server integration bugs and improving developer velocity
- Reduced read latency by **35%** through **Redis** caching for Firestore-based conversation history

Teaching Assistant — Jun. 2024 – Dec. 2024

Google X CIS 7000: Large Language Models | PyTorch, HuggingFace, RAG, LangChain — *Philadelphia, PA*

- Partnering with **Google** to migrate course infrastructure to **JAX** and **OpenXLA** for hardware-agnostic LLM assignments on TPUs, GPUs, and CPUs
- Created lecture slides and demos on reinforcement learning concepts, **BERT**, and attention mechanisms like **Flask Attention**
- Authored hands-on assignments on **Transformers**, **LoRA**, and **SliceGPT** for 50+ PhD students in a graduate LLM course

Software Engineering Intern — Jun. 2023 – Aug. 2023

Foxconn | Apache Spark, Tesseract, Python — *Taoyuan, TW*

- Reduced data storage costs by 25% by building **Spark** pipelines for de-duplication, compression, and aggregation
- Automated product data entry with a **Python OCR** tool, saving 50+ hours/month across teams

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, SQL, 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

Traffice Slice | *React, Express.js, SQLite, MitMProxy*

- Built a desktop privacy monitor that detects and flags suspicious data access by installed apps
- Intercepted requests via **TLS proxy (MitMProxy)**, filtered anomalies, and stored them in **SQLite** for real-time querying