

Mohammad Maaz Rashid

New York, USA — mr7374@nyu.edu — (929) 726-4505 — linkedin.com/in/rashid-maaz — github.com/codeholic08

SKILLS

- **Languages:** Python, C++, TypeScript, SQL, Bash
- **ML & GenAI:** PyTorch, Transformers, LLM fine-tuning, RAG, LLM evaluation, LangChain, ONNX Runtime
- **Retrieval & Data:** FAISS, HNSW, ChromaDB, PGVector, PostgreSQL, Redis, Spark, Hadoop
- **Cloud & Infrastructure:** AWS (S3, EC2, Lambda, EMR, Glue, SageMaker), Docker, Kubernetes, Git, Linux

EXPERIENCE

Qualcomm

San Diego, USA

Machine Learning Intern, GenAI

May 2025 – Aug 2025

- Fine-tuned **Qwen** and integrated a retrieval layer into the modern-framework QA pipeline to automate test-scenario validation across **9 internal test suites**, reducing manual validation time by **65%**.
- Developed an evaluation workflow for LLM-backed services measuring response quality, prompt efficiency, and redundant API calls, reducing LLM operating cost by **30%**.
- Architected a multi-agent orchestration layer coordinating six specialized workers, including layout, styling, component-generation, and validation, routed by confidence-based intent classification; **99% task completion** on an internal evaluation set.
- Engineered a distributed telemetry pipeline that ingested multiple gigabytes of daily device data into Amazon S3 and processed it on Amazon EMR using parallel Hadoop MapReduce jobs over HDFS, reducing telemetry-analysis turnaround time by **35%**.

Barclays

Pune, India

Software Engineer, AI & Backend Platforms

Aug 2022 – July 2024

- Fine-tuned FinBERT and Instructor-XL for financial-document classification and semantic retrieval, applying few-shot and post-training techniques and achieving **0.98 F1** on a held-out routing test set.
- Integrated ChromaDB-backed semantic retrieval into enterprise AI workflows supporting a platform credited with **eight-figure** business impact.
- Developed LLM-powered features—including chatbots, entity extraction, document search, and summarization—reducing manual workload by **45%**.
- Automated performance evaluation across **25,000+ customer records**, reducing end-to-end analysis time from multiple days to seconds.
- Engineered the FastAPI backend for a batch-pricing tool, processing thousands of user queries in parallel.

ACADEMIC EXPERIENCE

NYU Courant Institute of Mathematical Sciences

New York, USA

Graduate Research Assistant — Vector Databases and ANN Search

June 2026 – Present

- Developing **VecFast**, a high-update-rate vector database with a **two-tier ANN index**: an in-memory HNSW hot tier absorbing streaming inserts and a cluster-based cold tier, merged by an adaptive centroid algorithm that detects distribution shift.
- Benchmarking recall-latency tradeoffs against FAISS and HNSW baselines under sustained streaming inserts on NYU HPC.

NYU Courant Institute of Mathematical Sciences

New York, USA

Graduate Teaching Assistant — Computer Systems Organization

Aug 2025 – May 2026

- Led weekly recitations for **60+ graduate students** on CPU pipelines, memory hierarchy, virtual memory, and I/O systems; developed assessments with course faculty.

PROJECTS

ShaderLab: Generative GPU Image Editor — TypeScript, React, WebGL2, GLSL, Vite

[Demo]

- Developed an **LLM-assisted compiler and validation layer** converting natural-language edits into GLSL passes, rejecting shaders that fail compilation or color-correctness checks before rendering.
- Engineered a browser-based image editor on raw **WebGL2**, representing each adjustment as a reorderable GPU shader pass over a linear-space, premultiplied-alpha rendering pipeline.

Focus Assist: On-Device AI Pomodoro Tracker — Python, PyTorch, ONNX, Edge AI

[GitHub]

- Thread-safe Pomodoro timer for **Snapdragon X Elite** laptops with CLI task tracking and on-device **NPU inference** through an ONNX runtime path.

EDUCATION

New York University

New York, USA

M.S. in Computer Science

Sept 2024 – May 2026

- Coursework: Machine Learning, Deep Learning, Big Data, Algorithms, Foundations of Data Science

SRM Institute of Science and Technology

Chennai, India

B.Tech in Computer Science and Engineering

May 2022

AWARDS & LEADERSHIP

- **Winner, Qualcomm HaQathon** — on-device Snapdragon X Elite NPU application.
- **Vice President of Communications, NYU Graduate Student Council** — elected representative for 6,000+ students.
- **Barclays World Class Customer Service Award**