GAURANK MAHESHWARI

Rochester, NY | +1 (585) 957-6312 | gm8189@g.rit.edu | LinkedIn | GitHub | Portfolio

SUMMARY

Machine Learning / AI Engineer with 2+ years of experience specializing in LLM inference optimization, GPU acceleration, and low-latency model deployment. AAAI-26 Spotlight Author (Top 11%) and NSF Trainee focused on building production-grade, efficient systems (CUDA, TensorRT, vLLM). Pursuing M.S. in Data Science at RIT (Dec 2025). STEM OPT eligible.

PROFESSIONAL EXPERIENCE

Automation Engineer — Rochester, NY
Venture Creations Incubator (RIT) — Aug 2025 - Present

- Migrated 6+ internal workflows to unified database systems using Python + SQL ETL pipelines, improving data reliability and enabling real-time reporting across incubator operations.

AI Engineer — Rochester, NY
Magic Spell Studios — Jan 2025 - Aug 2025

- Designed LangChain multi-agent pipelines for podcast transcription (Deepgram ASR, Hugging Face, GPT); improved token usage to cut inference cost by ~35% and achieve ~86% speaker attribution accuracy on 20k+ episodes (~50k hrs).
- Engineered FastAPI APIs powering the MVP, deployed via AWS + GitLab CI/CD; added latency and cost monitoring that reduced iteration cycles by ~20%.

Applied Research Engineer — Rochester, NY
AWARE-AI NSF Research Traineeship Program, RIT — Aug 2024 - Aug 2025

- Optimized multimodal signal pipelines (EEG, EMG, ECG, motion capture) in PyTorch with CUDA (Python); reduced inference latency by ~30%, enabling near real-time deployment in human–robot collaboration (HRC) tasks.
- Benchmarked and accelerated ML inference with TensorRT vs. PyTorch baselines; improved throughput by ~1.5× and validated deployment on UR-10 and Sawyer robots for synchronized HRC experiments.

Machine Learning Engineer — Bengaluru, India
DeMons (18-employee venture-backed startup) — Aug 2022 - Jul 2023

- Built ML-driven scarcity and drop-timing models using Python/SQL, pandas, NumPy, scikit-learn; applied feature engineering to optimize transaction data, boosting launch engagement by ~18%.
- Revamped statistical valuation models for 600+ digital assets; automated release-timing recommendations and dashboards, improving market value perception of NFT assets by ~10% and strengthening early-stage trading liquidity.

PROJECTS

LLM Inference Lab: Research-grade Speculative Decoding Runtime

- Developed PyTorch + CUDA/MPS runtime with custom kernels, KV-cache reuse, batching, and CUDA Graphs, delivering ~2× GPU throughput, full reproducibility, and deterministic profiling; extended with multi-GPU optimizations.

CUDA Graph Routing: GPU-Accelerated Shortest Path on NYC Road Network

- Designed a CUDA-parallelized Dijkstra's algorithm on a 11.7M-node, 25.3M-edge graph using CSR layout and optimized SSSP kernels, achieving ~1.8× CPU speedup with memory-efficient loaders and scalable validation tools.

HONORS & AWARDS

- **AAAI-26 Spotlight (Top 11%): NewsLensAI** · Selected for Oral + Best Abstract shortlist; awarded travel funding from NSF AWARE-AI and RIT Graduate School to present in Singapore.
- **NSF AWARE-AI Research Trainee (2024–25):** Focus on transparent, multimodal ML and human-centered AI at RIT.

EDUCATION

Rochester Institute of Technology (RIT) — Rochester, NY
Master of Science in Data Science — Aug 2023 - Dec 2025

The LNM Institute of Information Technology (LNMIIT) — Jaipur, India
Bachelor of Technology in Computer Science — July 2019 - May 2023

TECHNICAL SKILLS

Languages: Python, C++, SQL
ML & LLM Systems: PyTorch, Hugging Face Transformers, LangChain, vLLM, Prompt Engineering, DeepSpeed
GPU & Acceleration: CUDA, TensorRT, Mixed-Precision, GPU Profiling, Parallelization, Quantization, Pruning
MLOps & Deployment: Docker, FastAPI, MLflow, AWS (EC2, S3, SageMaker, Bedrock), GitLab CI/CD, Kubernetes (K8s)
Data & Feature Systems: pandas, NumPy, FAISS, Redis, Pinecone
Foundations: Algorithms & Data Structures, OOP, Software Engineering Best Practices