SAMAY U. SHETTY

☎ +1 (585) 230-7406 — ✉ samayshetty.2025@gmail.com — linkedin.com/in/samay-shetty — samay-shettyyyy.github.io/portfolio

CS Grad Student and AI/ML Engineer with production experience leading teams to build RAG-based agentic systems and research expertise in disagreement learning (published at ACL). Shipped LLM-powered financial assistant serving 500+ daily queries and published research at ACL. Seeking Summer 2026 internship/co-op to innovate in both research and production

EDUCATION

Rochester Institute of Technology, Rochester, NY — Expected Dec 2026
Master of Science, Computer Science — GPA: 3.8/4.0
Relevant Coursework: Applications of Generative AI, Quantum Machine Learning, Machine Learning, Big Data Analytics

University of Mumbai, Mumbai, India — Aug 2024
Bachelor of Engineering, Electronics & Telecommunication — Minor in Data Science — GPA: 9.4/10

TECHNICAL SKILLS

ML/AI:: NLP, Deep Learning, LLMs, RAG,PyTorch, TensorFlow,
Languages: Python, C++, Java, SQL
MLOps: Docker, Kubernetes, AWS (EC2, S3, Lambda), FastAPI, Redis, CI/CD (GitHub Actions), WandB
Data & Vector Stores: ChromaDB, Pinecone, MongoDB, PostgreSQL, ETL Pipelines

EXPERIENCE

AI Engineering Intern — *Creating Wings Co. – Remote, Connecticut* — Nov 2025 – Present

- **Led a team of 5 in developing a personalized AI financial assistant helping women navigate finance, retirement, and education decisions**, serving 500+ daily queries with privacy-first user metadata collection for contextual, tailored guidance through FastAPI backend and React frontend (99.5% uptime)
- Optimized semantic search pipeline using ChromaDB vector store and custom retrieval strategies, achieving **66% latency reduction** (3.2s → 1.1s) and **35% improvement** in answer relevance through hybrid search and reranking
- Deployed scalable infrastructure on AWS (EC2 + ALB) with Redis caching layer, OAuth2 authentication, and monitoring dashboards, reducing cold-start latency by 40%

ML Research Assistant — *Lab of Population Intelligence,RIT – Rochester, NY* — May 2025 – Aug 2025

- **Advanced disagreement-aware NLP by enhancing DisCo neural architecture to capture diverse annotator perspectives**, integrating annotator metadata embeddings with cross attention and custom loss reweighting to improve label distribution predictions by **39%** across 4 benchmark datasets
- Optimized distributed training workflows on GPU clusters, reducing computational costs by **45%** through profiling, bottleneck analysis, and resource allocation strategies
- Co-authored paper accepted to **ACL NLPerspectives Workshop 2025** (presented at EMNLP), open-sourcing reproducible codebase (github.com/Homan-Lab/lewidi3) and contributing to Disagreement Learning research

PROJECTS

Savora AI – Restaurant Operations Intelligence Platform — *Python, Graph RAG, LLMs, SQLite*
github.com/Samay-Shettyyyy/Savora_AI

- **Eliminated manual spreadsheet analysis for restaurant managers through AI-powered natural language queries over sales, inventory, and scheduling data**, validated with 10+ stakeholder interviews across US and Asia
- Built agentic RAG system connecting disparate data sources using knowledge graphs and LLM orchestration, achieving 92% query accuracy with sub-200ms response times
- Designed multi-modal ETL pipeline ingesting 10K+ daily transactions into normalized SQLite warehouse

MatSAR – Polarimetric SAR Image Classification System — *C++, MATLAB, Computer Vision, CUDA*
Demo Video

- **Extended industry-standard PolSARPro software by adding ML-based terrain classification capability**, achieving superior accuracy through custom feature extraction from polarimetric decompositions of satellite imagery
- Accelerated inference pipeline with CUDA GPU optimization in MATLAB, achieving **35% reduction** in processing time for multi-gigabyte satellite image classification tasks

PUBLICATIONS

- **Sawkar, M., Shetty, S.U.**, et al. (2025). "LPI-RIT at LeWiDi-2025: Improving Distributional Predictions via Metadata and Loss Reweighting with DisCo." *ACL NLPerspectives Workshop (EMNLP 2025)*
- **Shetty, S.U.**, et al. (2025). "MatSAR: A Comprehensive Machine Learning Approach for PolSAR Data Processing." *International Journal of Computer Applications*