SAMBHAV SHRESTHA

sambhavshrestha111@gmail.com | New York, NY, USA | linkedin.com/in/sambhav101 | sambhavshrestha.com

EDUCATION

Stony Brook University - SUNY August 2025 - June 2027
Master's, Computer Science *GPA: 3.5*
Computer Vision, Natural Language Processing, Quantum Computing, Computer Architecture & Assembly, Operating Systems

St. Joseph's College New York September 2018 - June 2022
Bachelor's, Computer Science and Mathematics *GPA: 3.93*
Advanced Algorithms & Programming, Advanced Databases, Multivariable Calculus, Linear Algebra, Probability & Statistics

PROFESSIONAL EXPERIENCE

HCL Technologies March 2024 - July 2025
ML Infra Engineer at Meta *New York, NY, USA*
- Architected and deployed **Sherlock**, a production agentic AI system built with fine-tuned LLaMA and RAG to autonomously resolve ML infrastructure queries, reducing escalation volume and accelerating onboarding
- Engineered monitoring systems and a doctor tool using PyTorch to track production LLM health, setting up real-time alerting for entropy explosions, weight deviations, and throughput degradation
- Delivered end-to-end infrastructure support across the full experimentation and model deployment lifecycle, spanning CUDA kernels, distributed training, and model optimization

Tarifica June 2023 - February 2024
Software/Data Engineer *New York, NY, USA*
- Built 300+ web scrapers using Python and BeautifulSoup and designed Flask-based RESTful ETL pipelines to ingest, transform, and route structured data to downstream PostgreSQL analytics systems with strict latency requirements

Amazon July 2022 - March 2023
Software Development Engineer *Seattle, WA, USA*
- Designed and operated mission-critical microservices on AWS (EC2, Lambda, CloudWatch) powering Amazon Go and Just Walk Out retail stores
- Led migration of a high-traffic service to AWS, reducing costs, latency, and support tickets while improving reliability through automated CI/CD pipelines
- Delivered production features across Java, Kotlin, Python, TypeScript, and Ruby in a Linux environment

Microsoft June 2021 - July 2021
Data Science Student Research Fellow *New York, NY, USA*
- Extended the Financial Times police complaints study using R, uncovering novel race and gender patterns across datasets from NYC, Chicago, and Philadelphia
- Built regression and ML models to surface insights, presenting findings through compelling visualizations to data scientists and engineers at Microsoft Research

RESEARCH & PROJECTS

Sherlock — Agentic AI System at Meta *Python, LLaMA, RAG, PyTorch*
- Built a production Agentic AI system fine-tuned on Meta's internal Workplace forum data, combining RAG with agentic reasoning to autonomously resolve ML infrastructure queries for engineers and research scientists

Argument Quality Ranking: LLMs vs. Fine-Tuned RoBERTa *PyTorch, Hugging Face, scikit-learn*
- Benchmarked fine-tuned RoBERTa against GPT-5.5, Llama 3, and Mistral on pairwise argument quality ranking; developed RoBERTa v3 with margin ranking loss and test-time pair flipping, achieving 0.657 accuracy matching GPT-5.5 (0.665) at a fraction of cost; released on Hugging Face

Frequency Prior for Autoregressive Image Generation *PyTorch, CUDA, Lightning, wandb*
- Researched replacing attention-based transformers with Fourier frequency priors to improve stability, convergence, and visual quality in autoregressive image generation

Raft / Paxos Consensus Simulator *Rust, Python, TOML*
- Simulated distributed consensus with configurable node counts, network partitions, and leader election; visualized log replication and fault-tolerance under failure scenarios

SKILLS

Languages: Python, PyTorch, Java, Rust, C/C++, Kotlin, TypeScript, R, SQL, Verilog
ML / AI: Transformers, Computer Vision, LLM, CUDA, RAG, TensorFlow, OpenCV, scikit-learn
Cloud & Tools: AWS, GCP, Docker, Flask, PostgreSQL, Linux/Unix