BHANUJA KARUMURU

New York, USA | +1 3474459258 | linkedin.com/in/bhanujakarumuru | github.com/Bhanuu01 | bk3170@nyu.edu

EDUCATION

New York University, MS in Computer Engineering — New York, USA | Expected May 2027
Relevant Coursework: High Performance Machine Learning, MLOps, Deep Learning, Data Center & Cloud Computing

NIT Sikkim, B.Tech in Electronics and Communication Engineering — Sikkim, India | Dec 2020 – May 2024
Relevant Coursework: Machine Learning, Neural Networks, Deep Learning, Signal Processing, Digital Communications

WORK EXPERIENCE

Incoming Software Development Intern - Amazon.com Inc. — *Seattle, USA | Starting June 2026*

Software Development Engineer, SalesUp - Yellowlake Technologies Services Private Limited — *Kolkata, India | July 2024 – July 2025*
(Promoted from Software Engineering Intern)

- Built Flask-based microservices for a CRM platform, improving throughput by 40% through asynchronous task pipelines, REST API refactoring, and database optimization.
- Developed an LLM-powered email campaign feature using RAG and vector embeddings, increasing user engagement by 25%.
- Containerized applications with Docker and automated Jenkins deployments in AWS/Linux environments, reducing release overhead by 30%.
- Profiled and restructured async request handling, reducing P99 API latency by 25%.

Open-Source Contributor — *Github, Remote | Jul 2025 – Present*

- Merged upstream PRs in Conda (Python package manager) improving CLI error handling and dependency resolution; changes adopted by core maintainers and shipped to production.

RESEARCH EXPERIENCE

Student Researcher (TREx Fellow) - NYU Tandon School of Engineering — *New York, USA | Starting June 2026*

- Developing ML/DL-based computer vision pipelines for object detection, feature extraction, and geotagging from high-resolution historic maps, enabling GIS-compatible spatial data integration for urban infrastructure analysis under Prof. Debra Laefer.

AI/ML Infrastructure & Computer Vision Researcher, NYU RoboMaster (Team Ultraviolet) — *New York, USA | Jan 2026 – Present*

- Built and integrated real-time object detection and tracking pipelines for autonomous robots in NYU's VIP RoboMaster program, supporting vision model training and on-robot deployment with controls and hardware teams.

PROJECTS

FusedLinearAttention - Custom CUDA Kernel for Transformer Inference (github)

- Designed tiling strategy, bank-conflict-free shared-memory layout, and HBM traffic model for a three-kernel CUDA family fusing QKV projection and attention on H100 GPUs; 11/11 correctness tests pass with max absolute error $\leq 1.5 \times 10^{-7}$.
- Warp-cooperative bf16 hybrid achieved 1.22×–1.41× speedup over PyTorch baseline at short sequence lengths; estimated HBM reads reduced by up to 54.6%, peak GPU allocation cut 85–91%.

Intelligent Deadline & Expiry Detection System - Production MLOps

- Fine-tuned BERT NER (F1=1.0 on CUAD temporal expression subset) and RoBERTa classifier (F1=0.87) on CUAD legal dataset; built an end-to-end deadline extraction pipeline integrating both models into Paperless-ngx with continuous retraining triggered by feedback accumulation.
- Deployed on Chameleon with an asynchronous queue absorbing traffic bursts, auto-scaling to a second worker node on sustained queue depth, and production monitoring of inference latency, event creation volume, and correction rates via MLflow and Ray Tune.

SVG Generation - LoRA Fine-Tuning for Structured Code Generation

- Fine-tuned Qwen2.5-Coder-3B-Instruct with LoRA (Unsloth/TRL) on H100 for structured SVG output; iterated over 22 experiments on decoding strategy, XML repair logic, and repetition penalty, achieving competition score 0.93762 with 100% valid XML across 915 unique outputs.

StyleSync - Fashion Retrieval Recommendation System

- Built a two-tower deep retrieval recommender using TensorFlow Recommenders to learn user–item embeddings from interaction data, with FAISS-based ANN indexing pipelines achieving sub-100 ms end-to-end inference latency.
- Productionized the system with containerized FastAPI services, scalable offline training pipelines, model versioning, and health checks, enabling reliable rollout of retrained models.

TECHNICAL SKILLS

Languages & Systems: Python, C/C++, Java, Matlab, SQL, Bash, Linux, CUDA, Git
ML & Data Science: PyTorch, TensorFlow, Sklearn, HuggingFace Transformers, FAISS, XGBoost, PEFT (LoRA/QLoRA), RayTune, Pandas, NumPy
Generative AI & LLMs: OpenAI/GPT APIs, RAG Pipelines, Vector Databases, Prompt Engineering
ML Systems & Infrastructure: Docker, Kubernetes, MLflow, Jenkins, NSight Systems, Model Serving, Monitoring, CI/CD
Backend & Cloud: FastAPI, Flask, REST APIs, AWS (EC2, S3, Lambda)

PUBLICATIONS

Fusion of Data Augmentation for Improved Dysarthria Severity Classification, Journal of Signal Processing Systems, Springer, Accepted 2026

In-Domain Data Augmentation for Dysarthria Severity Classification, SPCOM 2024