Mrityunjay Bhanja

: /mewtyunjay

: mrityunjay.b@nyu.edu

: /in/mewtyunjay

Education

New York University Aug 2023 – May 2025(Expected)

Master of Science in Computer Science *New York, NY*

Coursework: VR/AR, Computer Vision, Machine Learning, Algorithms, Big Data, AI, Human Computer Interaction

Amity University Aug. 2017 – May 2021

Bachelor of Technology in Computer Science and Engineering *Gurugram, India*

Experience

NYU - CREATE Lab Apr 2024 – Present

Machine Learning Engineer *New York, NY*

- Built real-time multi-agent system with RAG pipelines and 5 OSS LLMs (Llama, Qwen, Deepseek) as an adaptive learning platform for K-12 students; implemented cross-agent reasoning for educational scaffold verification, improving learning objective alignment by 30% and serving 1000+ students with 93% satisfaction.
- Architected and implemented end-to-end ETL pipelines leveraging AWS Lambda, S3, and MongoDB Atlas to efficiently process and analyze 14,000+ images using InternVL2.5 on AWS SageMaker.

Maersk Aug 2021 – Aug 2023

Machine Learning Engineer *Bengaluru, India*

- Engineered Apache Spark observability system (monitoring, lineage, config-driven DAGs) as a critical component of DataLab - an in-house Databricks alternative that drove $4M annual savings
- Engineered and optimized a data processing pipeline using **PySpark** to handle 155GB of data for a freight cost estimation model, improving data processing efficiency and enhancing cost estimation accuracy by **20%.**
- Engineered end-to-end data lineage system in Python, integrating MongoDB and Azure Databricks to unify Maersk's data engineering workflows; reduced failure detection time by 40%, saving 520 engineering hours annually.
- Administered a voice analytics systems using **OpenAI's Whisper**, for real-time transcription of phone calls, enabling insights and reduced turnaround time. Achieved a **word error rate of 30%** on transcribed phone calls.

Hewlett Packard Enterprise April 2021 – Jul 2021

Machine Learning Engineer Intern *Bengaluru, India*

- Built end-to-end traffic violation system using ONNX-optimized fine-tuned YOLOv3 and SSDMobilenetV2 (0.86 mAP); deployed on HPE edge hardware with automated license plate OCR for real-time violation tracking
- Implemented **HoloGAN** to create a 3D image generator capable of reconstructing 3D facial images from a single 2D photograph with low latency on Azure Cloud.

Skills

Languages: Python, C++, SQL, JavaScript, Go, GraphQL

Machine Learning: PyTorch, TensorFlow, NLP, MLX, CoreML, MLOps, YOLO, OpenCV

Big Data & Cloud: Apache Spark, Kafka, AWS (S3, EC2), Azure, Docker, Kubernetes, Airflow, HPC

Databases: MongoDB, MySQL, PostgreSQL, Redis, Pinecone, Weaviate, ChromaDB, DynamoDB

Tools & Frameworks: Langchain, LlamaIndex, Diffusion, FastAPI, Flask, Git, CI/CD

Projects

Multi-Agent Voice Assistant: Built on-device multi-agent voice assistant for programming assistance, creative ideation, and therapy support using MLX-optimized quantized LLMs (Llama, Qwen, Gemma) with LoRA fine-tuning; engineered agent orchestration achieving real-time speech-to-speech latency.

3D Try-On in AR: Implemented an iOS AR virtual try-on system (Unity AR Foundation, ARKit, C#) with real-time 3D body tracking, enabling seamless clothing visualization through custom prefab controllers and skeleton mapping.

Scaling Test Time Compute for local SLMs: Implemented test-time compute scaling with MLX on GPU, using beam search, Process Reward Model, and reinforcement learning based optimization to achieve 45% accuracy on MATH benchmark with 3B parameter open source model - matching 70B model performance on device, locally

Football Analysis Using CV: Engineered a real-time football analysis system using YOLOv11x, ByteTrack, and YOLOv-pose models to generate tactical visualizations through homography-based mapping.

Publications

Model Selection for Parkinson's Disease Classification Using Vocal Features: Bhanja, M. Chaudhary, S.