Anuj Patel

New York, NY | 347-401-7880 | amp10162@nyu.edu | www.panuj.com | linkedin.com/in/panuj

EDUCATION

New York University **Sep 2023 – May 2025**
Master of Science in Electrical Engineering **(GPA: 3.9/4.0)** New York, NY

- **Co-authored** a textbook on **"Fundamentals of Communication Theory"** with **Dr. Unnikrishna Pillai**.
- Coursework: ML, Deep Learning, CV, High Performance Machine Learning, Probability, Big Data
- Research: Developed a mmWave channel sounder at 57.51 GHz under **Dr. Sundeep Rangan** for wireless channel measurements.

Vellore Institute of Technology **Jul 2019 – May 2023**
Bachelor of Technology in Electronics and Communication Engineering **(GPA: 9.2/10, Rank: 4)** Vellore, India

- Coursework: Applied Linear Algebra, Statistics, Cryptography and Network Security, Object Oriented Programming

SKILLS

Languages	Python, TypeScript, C/C++, Golang, CUDA, MATLAB, SQL, Bash
Frameworks	PyTorch, HuggingFace, LangChain, LangGraph, GraphQL, Numpy, Pandas, Wandb
Cloud	AWS (SageMaker, EC2, ELB, S3, Redshift), GCP (Vertex AI, BigQuery, AutoML)
DevOps	Kubeflow, Airflow, Spark, Kafka, Kubernetes, Docker, CI/CD, Git, Slurm
Databases	PostgreSQL, MongoDB, Weaviate, Pinecone

WORK EXPERIENCE

New York University **Jan 2024 – Dec 2024**
Machine Learning Engineer New York, NY

- Adapted and deployed a **RAG-based GenAI assistant** using **LangChain**, **OpenAI APIs**, and **Pinecone**, enabling natural language search over 100+ academic and policy documents.
- Engineered modular retrieval pipelines with **text-embedding-ada-002**, **Docker**, and **GitHub Actions**, achieving **<300ms median latency** and readiness for seamless internal rollout.
- Developed and productionized a personalized course recommendation engine using **TF-IDF**, **cosine similarity**, and **user embeddings**; served via **FastAPI + PostgreSQL** for 5K+ students.

Johnson & Johnson **Jun 2024 – Aug 2024**
Data Science Intern New Brunswick, NJ

- Led development of an **LLM-powered clinical assistant** using **LLaMA 2–7B** with **LoRA fine-tuning** over 10M+ anonymized patient records—reducing physician query time by **46%** and influenced $12M+ in operational savings.
- Built and productionized a **real-time, multimodal ML pipeline** (text, imaging, vitals) using **PyTorch, HuggingFace and AWS**—achieved **28% lift in outcome prediction accuracy** and scaled to serve 20M+ patient records.

Indian Space Research Organization **Dec 2022 – May 2023**
Machine Learning Researcher Ahmedabad, India

- Trained and deployed a **GAN-based super-resolution model** on RISAT-1A SAR data, boosting spatial resolution **2×** while retaining speckle-aware texture priors for terrain analysis.
- Integrated and optimized **quantized CNNs** in an Edge AI framework for real-time cloud detection in Microsat's onboard inference pipeline, reducing transmission data by **67%** under compute and power limits.

PROJECTS

Efficient Federated Learning using Gradient Pruning and Adaptive Methods | *PyTorch* **Sep 2024 - Dec 2024**

- Pioneered an efficient FL framework with gradient pruning and adaptive federated optimization, **reducing training time by 22%** and boosting model generalization.
- Increased **bandwidth efficiency by 143%** via gradient compression and mix-precision training, validating ResNet accuracy with **PyTorch DDP, DeepSpeed and Hugging Face Accelerate.**

Transformer-Based Multi-Modal Emotion Recognition System | *PyTorch, OpenCV, HPC* **Sep 2024 - Dec 2024**

- Enhanced a transformer based framework for emotion recognition, **achieving 33.96% top-1 and 98.13% top-5 precision** on RAVDESS data integrating both facial and vocal cues.
- Applied **advanced modality fusion techniques** with feature extraction and preprocessing pipelines, **processing 4,000+ video and audio signals to improve robustness** in noisy/incomplete datasets.

Movie Recommendation System with NCF | *Python, PyTorch, PySpark, SQL* **Sep 2024 - Dec 2024**

- Architected a scalable movie recommender system using **Neural Collaborative Filtering (NCF)**, achieving **52% Hit Ratio** on MovieLens 1M via distributed preprocessing using **Apache Spark** and SQL-based warehousing.
- Redesigned a 1M-record ML pipeline with optimized feature engineering, negative sampling, and a SQL backend—**reducing retrieval latency by 34%** and integrating seamlessly with a Streamlit interface.