Anuj Patel

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

WORK EXPERIENCE

Pointel **Feb 2025 – Present**
AI Engineer New York, NY

- Remodeled multi-agent orchestration layer with **LangGraph** and **AWS Bedrock**, enabling planner–executor agents to autonomously chain API workflows and reduce human intervention to just **3 hours/week**.
- Improved RAG quality with **ChromaDB** hybrid search (dense + BM25), tuning chunking/scoring to acheive **Precision@10 of 0.84** and **sub-200ms latency** at 1K+ document scale.
- Collaborated with business users to identify pain points and **develop better KPIs for agent evaluation** strategies and improved satisfaction rate.

New York University **Jan 2024 – Dec 2024**
AI Engineer New York, NY

- Productionized a **RAG-powered chatbot** using **LangChain**, **OpenAI APIs**, and **Pinecone**, enabling natural language search over 100+ academic and policy documents with **sub-300ms** median latency .
- Engineered **retrieval and recommendation pipelines** using **text-embedding-ada-002**, **FastAPI**, and **PostgreSQL** in a **Docker CI/CD** environment, powering course recommendations for **5K+ students**.

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

- Led development of an clinical assistant using **LLaMA 2–7B** with **QLoRA and PEFT** over 1M+ anonymized patient records; reducing query time by **46%** across **450+ daily clinical queries** during pilot phase.
- Built a **real-time, multimodal ML pipeline** (text, imaging, vitals) using **PyTorch, HuggingFace** and **AWS**; achieving **28% lift** in outcome prediction accuracy and scaling to 20M+ patient records.

Indian Space Research Organization **Dec 2022 – May 2023**
Artificial Intelligence Researcher Ahmedabad, India

- Trained a **GAN-based super-resolution model** on SAR data, **doubling** resolution while preserving details for terrain mapping.
- Integrated **quantized CNNs** for real-time cloud detection, cutting transmission data by **67%** under power constraints.

EDUCATION

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

- **Co-authored** an engineering 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

PROJECTS

Multi-Agent RAG Chatbot for E-commerce Customer Support **Aug 2025 – Sep 2025**

- Designed an **agentic RAG chatbot** with **7 tools** and **Voyage-3.5 embeddings**, handling **20K+ parts** via dual-database orchestration (SQLite FTS5 + ChromaDB) in **2 LLM calls**.
- Implemented a **multi-agent workflow** with RAGOrchestrator for **intent classification** and **semantic retrieval**, enabling real-time troubleshooting and parts compatibility.

Efficient Federated Learning using Gradient Pruning and Adaptive Methods **Sep 2024 – Dec 2024**

- Pioneered a FL framework that **reduced training time by 22%** and boosted generalization using adaptive optimization.
- Achieved **143% bandwidth efficiency gain** via gradient compression and mixed-precision training with **PyTorch DDP, DeepSpeed, and HF Accelerate**.

Transformer-Based Multi-Modal Emotion Recognition System **Sep 2024 – Dec 2024**

- Enhanced a transformer for emotion recognition, achieving **33.9% top-1** and **98.1% top-5 precision** on the RAVDESS dataset.
- Applied advanced **fusion techniques** over **4K+ video-audio samples**, improving robustness in noisy and incomplete datasets.

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, SQLite, ChromaDB, Weaviate, Pinecone