# REET NANDY

reet.nandy@nyu.edu | reetnandy.com | github.com/techpertz | linkedin.com/in/reetnandy | +1(518)9306116 | Willing to Relocate

## SUMMARY

Full-Stack AI Engineer with **3** years of experience across 6 internships. Skilled in scalable microservices, event-driven architectures and AWS deployments. Currently **building projects** involving **multi-modal AI agents** with LLMs, vision models, and RAG pipeline.

## SKILLS

**Frontend + Backend:** Python (FastAPI, Django), TypeScript (Next.js), Tailwind CSS, Node.js (Express.js), Java (Spring Boot)
**AI / LLM:** Transformer (GPT, BERT, LLaMA), Graph + RAG, Fine-tuning, Embeddings, Quantization, Multi-Agent
**Cloud + DevOps:** AWS (EC2, Lambda, S3, SageMaker, Prometheus), GCP, Kubernetes, Docker, Jaeger, CI/CD, ELK, Microservices
**Data + Pipeline:** PostgreSQL, MongoDB, Vector DBs, Redis, Kafka, Airflow, ETL / ELT, REST, gRPC, Agile

## PROFESSIONAL EXPERIENCE

**Mobility Intelligence** New York City, USA
*Backend Development Intern (Machine Learning)* June 2024 – December 2024

- Built a real-time price prediction system using regression and Kalman filtering, achieving less than 5% error on 90-day forecasts.
- Integrated live model predictions into a React frontend with sub-1s latency sustaining 150k+ daily requests with 99.9% uptime.
- Automated ETL pipelines in Airflow to process 15M+ daily records, reducing manual effort and enabling downstream analytics.
- Set up Prometheus + Grafana with SLIs and alerting, cutting MTTD by 60% and boosting incident response efficiency.
- Deployed ML-backed FastAPI microservices with Celery, Redis, and autoscaling on AWS EKS along Helm charts.

**Defence Research & Development Organisation** India
*Software Engineering Intern (R&D)* January 2023 – June 2023

- Engineered multithreaded architecture for real-time LiDAR processing, handling 50K data points/sec (97% accuracy).
- Implemented Redis-based geospatial caching over PostgreSQL/PostGIS, reducing GPS query latency from 1000ms to 150ms.
- Developed ETL pipeline using memory-efficient streaming, processing 12GB/min while reducing memory usage by 60%.

**Solar Industries India Ltd** India
*Software Engineering Intern* April 2022 – December 2022

- Led a team of 5 to automate workflows, delivering 5 production ready Django systems that standardized 80% of manual processes.
- Reduced API latency by 25% and integrated distributed tracing with Jaeger and Grafana, enabling real-time debugging.
- Designed a partitioned Kafka pipeline with scalable consumer groups, processing 2.5M+ rows/sec using Redis and Cassandra.

## PROJECTS

*[AI Agent]* **GraphRAG - LLM Document Compliance** (Github) *April 2025*

- Launched an Agentic SaaS with real-time document edit, approval and audit reports via Graph based RAG and LLM.
- Implemented a GraphRAG and PDF parser from scratch for unstructured PDFs using PDFMiner & Tesseract (OCR).
- Executed semantic chunking + NLP NER + BART-CNN summarization, achieving 70% relation extraction accuracy.
- Synthesized hybrid retrieval (Vector + Graph + metadata) boosting compliance accuracy to 90%.

*[Python / Core]* **Hierarchical Vector Database from Scratch** (Github) March 2025

- Built embedding database (library - document - chunk) with async collection mutexes; 12K ops/sec at <0.1% conflicts.
- Added 3 indexing algorithms (LinearScan/KD-Tree/LSH) for vector search on 10M vectors in 18ms.
- Led Kubernetes Helm deployment along with custom made CLI toolkit, reducing onboarding complexity by 100%.

*[AI Workflow]* **Multi-Platform LLM Task Orchestrator** (Github) December 2024

- Created Flask API with Gmail/Slack/Trello via OAuth2, automating task extraction and priority classification with OpenAI.
- Configured MongoDB RAG system with vector indexing for context memory management via chat history tracking.
- Automated cross-platform workflow triggers by combining LLM with rule-based logic, reducing manual task handling by 60%.

*[AWS / Fullstack]* **AI-Fitness Analytics Dashboard** (Github) April 2024

- Deployed a Django API on AWS Elastic Beanstalk with Google Fit integration, leveraging auto-scaling and health checks.
- Built pipelines using Lambda and SageMaker to deploy a KNN model and an SNS/SQS ingestion achieving 800ms P95 latency.

## EDUCATION

**New York University** – New York City, USA September 2023 - May 2025
***Master of Science, Computer Science (MS)*** | *Merit Scholarship Recipient* GPA: 3.7/4.0

- *Relevant Coursework*: Data Structures and Algorithms, Cloud Computing, Machine Learning, Big Data Analytics, Database Systems, Software Engineering, Object-oriented Design, Probability and Statistics
- *Graduate Teaching Assistant*: CS GY 6233 - Operating System, CSCI UA 0310 - Algorithms

**Manipal University Jaipur** – India July 2019 - May 2023
***Bachelor of Technology, Computer Science (B. Tech)*** GPA: 3.8/4.0