Ritesh Ojha

Brooklyn, New York 11220 | Phone: +13479407232

Email Id: ro2283@nyu.edu | **LinkedIn**: https://www.linkedin.com/in/ritzzi23/ | **Github**: https://github.com/ritzzi23

EDUCATION

New York University, New York City, NY, USA — Sep 2024 – May 2026

Master of Science in Computer Engineering — (GPA: 3.8/4)

Amity University, Noida, India — July 2016 – May 2020

Bachelor of Technology in Computer Science and Engineering — (GPA: 3.9/4)

WORK EXPERIENCE

The Warehouse Group | Gurugram, Haryana, India — Oct 2023 – Aug 2024

Data Scientist — **Ranking, Recommendation Systems, NLP, FastApi**

- Developed and implemented an advanced recommendation system with ranking, NLP, content recommendations, embedding, information retrieval using collaborative filtering, content-based filtering, and hybrid models with TensorFlow, Keras, and Python for E-commerce platform.
- Collaborated with the team to develop a sentiment analysis framework utilizing advanced NLP techniques such as text preprocessing, vectorization, and machine learning classification algorithms, achieving 87% accuracy in automating user comment classification while enhancing response times by 25%.

Accenture | Bengaluru, Karnataka, India — Oct 2020 – Sep 2023

Software Engineer — **Go, RabbitMQ, MongoDB, Docker, WebSockets**

- Developed an API to provide optimized routes between delivery locations using graph algorithms.
- Incorporated geospatial queries using PostGIS to calculate distances.
- Wrote integration tests for route validation and containerized the backend for Kubernetes deployment.

PROJECTS:

Embeddings & Vector DB Project: Smart Notes Linker — Jan 2025 – May 2025

- Built a context-aware note linking system using Sentence-BERT and OpenAI embeddings to capture semantic similarity between notes, Indexed note embeddings with FAISS for fast and efficient vector similarity search
- Developed dynamic suggestions of related notes upon editing or creating, using top-k vector similarity
- Integrated a real-time UI with preview and link insertion via Streamlit or VSCode plugin and Implemented filters for time relevance, tags, and custom weights to refine results.

Deep Learning Project — Jan 2025 – June 2025

- Implemented CNN, Bidirectional GRUs/LSTMs, transformers, Variational Encoders, GAN from scratch in PyTorch
- Designed and trained a **Vision Transformer (ViT)** for image classification, incorporating Patch Embedding, multi-head self-attention, transformer layers, residual connections, and layer normalization.
- Worked on a Joint Embedding Predictive Architecture (JEPA) using Energy-Based Models to build a self-supervised world model applied to a wall-maze environment.

LLM Projects: — September 2024 – May 2025

- BBuilt end-to-end Retrieval-Augmented Generation (RAG) applications using LangChain, implemented document ingestion, intelligent chunking, and semantic search.
- Integrated open-source LLMs with quantization techniques like LoRA, QLoRA to reduce memory footprint
- Deployed the application using FastAPI, exposing RESTful endpoints for scalable inference and downstream integration and enable local inference
- Exploring LangGraph for building multi-step LLM and AI agents with memory and tool-use capabilities

Go-Based Fleet Vehicle Monitoring Microservice — Jan 2024 – June 2024

- Developed a Kubernetes-deployed Go microservice to ingest real-time vehicle telemetry (speed, GPS, fuel), achieving scalable data collection from a simulated fleet of 1000+ nodes.
- Designed RESTful APIs for telemetry ingestion and analytical querying; used PostgreSQL for persistent storage and Redis for low-latency caching.
- Implemented infrastructure as code with Helm charts and instrumented the service with Prometheus and Grafana for live monitoring and alerting.

TECHNICAL SKILLS

- Python, Pytorch, MLFlow, HuggingFace Transformers, Langchain, LangGraph, Git, Docker, Optuna (Hyperparameter Tuning)