ROHIT NANJUNDAREDDY

+1(312) 539-2733 rnanj@uic.edu linkedin.com/in/rohit-nanjundareddy/ github.com/rohitnanjundareddy

Summary

Machine Learning / AI Engineer with 4+ years building production-grade ML and LLM systems across vision, language, and multimodal domains. Developed anomaly detection, chat intelligence, and recommendation engines, along with data-efficient pipelines using transfer learning, few-shot learning, and synthetic data generation. Experienced in LLM/RAG systems, Vision Transformers, and scalable ML infrastructure with Kubernetes, Docker, Spark, and EMR, focused on end-to-end automation and deployment.

Education

MS in computer science — University of Illinois Chicago **Chicago, IL**
Grade: 3.85/4 *August 2024 - May 2026*

Bachelors in Computer Science — Amrita School of Engineering **Bangalore, India**
Grade: 3.4/4 *2016 - 2020*

Technical Skills

ML & Deep Learning: PyTorch, TensorFlow, ViT, Transformers (T5), DQN, ResNet, Diffusion Models, OpenCV.
LLMs & RAG: OpenAI, Gemini, Bedrock, Ollama, LangChain, Vector DBs (Qdrant, FAISS)
Big Data & Cloud: Spark, Hadoop, EMR, GKE, Kubernetes, Docker.
Programming: Python, C++, Java, JavaScript.
MLOps & Tools: Airflow, MLflow, Terraform, Kafka, gRPC, ETL Pipelines
Databases: Neo4j, MongoDB, Redis, BigQuery, DynamoDB.
Visualization: D3.js, Chart.js, Matplotlib.

Experience

Graduate Research Assistant — University of Illinois Chicago **May 2025 - Present**

- Implemented multi-MLP gene expression model using Prov-GigaPath embeddings
- Developed a ViT-based breast cancer prediction model using mammograms with diffusion-based data augmentation.
- Built scalable histopathology pipelines handling 100k+ tiles.

Senior Data Scientist — 247.ai **May 2023 - August 2024**

- Built and deployed deep-learning anomaly & misuse detection (ResNet + Python), reducing QA review time 40%.
- Designed a chat-routing recommendation engine using agent-behavior embeddings, improving first-call resolution 25%.
- Engineered scalable ETL and ML pipelines using Docker, Kubernetes, Spark, speeding model iteration cycles 30%.

Data Scientist — 247.ai **July 2020 - April 2023**

- Developed sentiment analysis and speech-to-text models using CNN and BiLSTM architectures, achieving 92% and 94.2% accuracy through optimized feature extraction and sequence modeling.
- Developed a predictive workforce optimization system using SQL and ML models (Gradient Boosting + Time-Series Forecasting), improving productivity by 15% and reducing idle time by 10%.
- Created React/TypeScript dashboards enabling faster decision-making across support teams.
- Built NLP/RNN-based agent-feedback learning systems raising CSAT 20%.

Projects

Few-Shot & Transfer Learning for Medical Imaging – *PyTorch, Prototypical Networks, Diffusion Models* (**Aug-Dec 2025**)

- Developed a data-efficient abnormality detection pipeline on the CheXpert dataset using few-shot and transfer learning with CNNs (ResNet-18, DenseNet-121) and a medical-domain ViT (RAD-DINO) with Prototypical Networks classifier, achieving 50.5% accuracy in 10-shot settings through domain adaptation and diffusion-based augmentation.

Intelligent Wumpus World Agent – *POMDPs, Knowledge Graphs, LLM-Based Reasoning* (**Jan-May 2025**)

- Built a LLM-driven reasoning agent that uses a structured knowledge graph for environment understanding, achieving a score of 550/1000 through belief-state search and LLM-based action planning.

InsureEase (insure-ease-kappa.vercel.app) – *Sentence Transformers, RAG, Vector Databases, Neo4j* (**Jan-May 2025**)

- Developed an LLM-powered insurance document analysis system that uses RAG pipelines and a knowledge graph to compare policies, extract insights, and generate personalized recommendations.

Building LLM from Scratch – *JTokkit, AWS (EMR, S3, Lambda), Hadoop, Spark* (**Aug-Dec 2024**)

- Implemented an end-to-end LLM training workflow using Hadoop and AWS EMR, and built a scalable conversational agent by integrating Amazon Bedrock and Ollama through AWS Lambda and gRPC.

Active Research

- LLM-based multi-agent in game theory using reinforcement learning agent (GNN+RL+DQN)
- Predicting 5-year breast cancer risk from mammography images