ANJALI PATIL
Boston,MA | patil.anjali2023@gmail.com | +1 (857) 310-9395 | LinkedIn

Education

Northeastern University, Boston, USA - Master's in Analytics **(AI/ML)** | GPA:**(4.0/4.0)** **Dec 2025**
Coursework: Machine Learning | Deep Learning | Probability and Statistics | Geospatial Data Sciences

Narsimha Reddy Engineering College, JNTUH - B-Tech. in Computer Science | GPA: **(3.4/4.0)** **July 2023**
Coursework: Software Engineering | Data Science | Web development | Machine Learning | Data Mining

Experience

AI engineer intern | AMD | Colorado, United States **January 2025 – June 2025**

- Developed a **RAG** and Summarization Application using AMD's **Llama-3.2 model** and **Ryzen AI 1.4**, implementing hybrid CPU/NPU inference to improve latency and throughput**.**
- Accelerated Embedding Inference **3.5x** for large-scale vector retrieval (>**1M** documents) by offloading the **BGE Large v1.5 model** to the NPU, enabling efficient and scalable data processing.
- Co-authored AMD's foundational **RAG** templates for **Ryzen AI PCs**, creating a modular, production-ready framework to accelerate the development of AI-native applications for the developer community.
- ***Authored a featured technical blog on AMD's official website detailing implementation of RAG with Hybrid LLM on AMD Ryzen AI Processors.*** link

Data Strategy Intern | MedPlus Diagnostics| Hyderabad, India **April 2023 – June 2023**

- Migrated diagnostic scripts from SnowSQL to Databricks SQL, improving workflow runtime from 3 hours to 45 minutes, aligning with stakeholder goals and ensuring process efficiency.
- Implemented a Shortest Path algorithm with IoT-enabled monitoring systems, improving operational efficiency by 25% and ensuring 95% on-time sample delivery.
- Participated in formal experiments and contributed to the optimization of operational routes using AWS Fargate and Docker images. Provided insights on tool features to enhance operational models.

Projects

NANDA: Intelligent Multi-Agent Research Project (MIT & Akamai Collaboration) ***(Ongoing)***
Tech Stack: Python, FastAPI, AWS EC2, MongoDB, MCP, A2A, SLIM, Docker

- Working with MIT and Akamai on Project NANDA, developing an intelligent multi-agent framework leveraging MCP, A2A, and SLIM protocols for real-time communication, contextual reasoning, and autonomous task execution.
- Deploying modular agents on AWS EC2 with MongoDB telemetry, applying the system to enterprise **automation, adaptive routing, and trip planning** through REST API and real-time data integration.

Dynamic QnA Document Bot: *VectorRAG*, *LLM*, *OpenAI, Llama Index, Pinecone DB, Python*

- Developed a solution to reduce the resolution time for service and customer care personnel by creating a Chatbot that can parse in any number of documents and become a QnA machine on them.
- Created **Document chunks with metadata** for each, used **semantic similarity** to retrieve top 5 relevant chunks and finally ensembled **multi-query retrievals** as context for the **LLMs** using prompt engineering.

Extended-Reality Remote Assistance for Manufacturing with LLM-driven Avatars

- Engineered context-aware remote assistance avatars using **Extended Reality** and **Large-Language Models** with domain-specific prompt engineering, reducing manufacturing downtime by **35%.**
- Integrated **C# scripts** in **Unity3D** with speech-to-text, chat completion, and text-to-speech APIs, achieving **92%** accuracy in troubleshooting industrial equipment.

Technical skills

Programming Languages: Python, SQL, R, PySpark, No-SQL, Sklearn, TensorFlow, Pandas, NumPy
Machine Learning: Hypothesis Testing, Casual Inference, EDA, Prediction, Classification, NLP, LLM, GenAI
Cloud Services: Databricks, BigQuery**,** EC2, S3, IAM, DynamoDB, VPC, VertexAI, Looker
Packages: PyTorch, spaCy, AutoML , SciPy, seaborn, TensorFlow
Tools/Software: Kafka, Spark,Power BI,Qlik Sense, Docker, Apache Airflow, MS Excel, Tableau,Linux CLI.