Nithin Reddy Yanna

Buffalo, NY ☎ +1-716-907-8910 ✉ nyanna@buffalo.edu in https://www.linkedin.com/in/nithin-yanna-716054217

SUMMARY

Data Science graduate student with hands-on experience developing AI systems for forecasting, document intelligence, and LLM pipelines. Skilled in time-series modeling, NLP, and deploying research-grade models into production using PyTorch, LangChain, and MLOps best practices to build scalable, production-ready solutions while collaborating with cross-functional teams.

EDUCATION

University at Buffalo | *Masters in Engineering Science, Data Science* **Aug 2024 - Present**

Amrita Vishwavidya Peetham | *Bachelors in Computer Science and Engineering* **Oct 2020 -Apr 2024**

Work Experience

AriesView(Real Estate AI startup) | *AI Research and Engineering Intern* **May 2025 - Present**

- Built a retrieval-augmented using **LangChain**, **ChromaDB**, and **LLMs** (Gemma, GPT-4); forecasted a prototype combining lease data with macroeconomic signals using LLM agents and vector search.
- Evaluated and integrated **OCR tools** (Unstructured, Azure Vision, Tesseract) to improve document parsing accuracy across leases, tax bills, and deeds by over **35%**.
- Developed **document intelligence workflows**, including chunking, embedding and semantic search using **FAISS**.
- Benchmarked **LLM hallucination rates** and implemented prompt tuning and retrieval filters to reduce false data generation by ~40%
- Investigated use of **Azure AI Studio** and **LlamaIndex** as LangChain alternatives to optimize cost, and speed in production settings.
- Contributed to LLM evaluation and prompt design for legal Q&A, time-variant clause tracking, and anomaly detection in contract terms.

PROJECTS

LLM Agent Evaluation & Reasoning Analyzer **Feb 2025 - Present**
Personal Project

- Developed an agent-based evaluation framework for Android automation tasks using **Gemma 12B** via **Ollama**, simulating real-world tasks like messaging, uninstalling apps, and setting alarms.
- Implemented **few-shot prompting**, **chain-of-thought reasoning**, and a **self-reflection module** where the agent critiques its own actions to improve future decision-making.
- Designed evaluation tools measuring **step-level accuracy**, **episode success rate**, and **task-specific breakdowns** (e.g.,take_photo = 100%,uninstall_slack = 67%).
- Built a structured error analyzer for visualizing common mistakes (wrong element click, wrong action type) and identifying failure modes in LLM output.
- Integrated plotting and PDF report generation for benchmark visualization, supporting real-time debugging and model comparison.

Loan predictor **Nov 2024 - Dec 2024**
University at Buffalo

- Performed EDA and label encoding for categorical variables to prepare data for machine learning.
- Achieved 92% accuracy with XGBoost, outperforming models like Logistic Regression, Random Forest, and Gradient Boosting.
- Used DagsHub to track experiments, log metrics, and version datasets/models for reproducibility.
- Integrated the model into a FastAPI backend for real-time predictions, with a Streamlit interface deployed online for user access.

Taxi Demand Prediction **Feb 2025 - Mar 2025**
University at Buffalo

- Developed a machine learning pipeline for predicting taxi demand using Hopworks Feature Store and MLflow.
- Trained LightGBM models with hyperparameter tuning, improving MAE to under 6 rides per hour.
- Automated training and deployment using GitHub Actions, enabling continuous integration and deployment.
- Built an interactive Streamlit app to visualize taxi demand predictions and provide real-time insights.

NYC Taxi Dashboard Analytics **Mar 2025 -Apr 2025**
University at Buffalo

- Designed and implemented an end-to-end data analytics pipeline for NYC taxi data using AWS services.
- Automated data extraction and transformation with AWS Lambda and AWS Glue Crawler to structure raw data.
- Optimized data querying and aggregation in Amazon Athena for efficient analysis and reporting.
- Developed an interactive Power BI dashboard, integrating AWS data via Amazon ODBC drivers, enabling real-time insights into taxi demand, trip distribution, and fare trends.
- Ensured scalability and performance by leveraging cloud-based architecture for seamless data processing.

TECHNICAL SKILLS

- **Programming Languages**: Java, Python, C/C++, R
- **Data Science & Analysis**: Deep Learning, Natural Language Processing (NLP), Large Language Models (LLM), Gen AI, Data/ETL pipelines, LangChain pipeline, MLOps workflow, Data Analysis, Operations Research, Project Scoping, PGvector, FAISS, Deviate
- **Tools & Platforms**: JavaScript, HTML/CSS, jQuery, Data Structures and Algorithms, OOPs, Spring Boot, React, JUnit, WordPress, Material-UI, MySQL, MongoDB, AWS S3, pandas, NumPy, Matplotlib, PyTorch, Git, CI/CD, IntelliJ, Eclipse, Jupyter Notebook, Android Developer, Google BigQuery, Streamlit, SAP Dataspare, Database Systems, Spark, Dask, AWS, GCP, Azure, Ad Tech Knowledge

ACHIEVEMENTS AND CERTIFICATIONS

- **Google Cloud Certification** | 30 days of Google Cloud - Fundamentals