Rahul Yadav

+1(716)238-0835 | rahulyad@buffalo.edu | linkedin.com/in/rahul-yadav-704 | Github | Website

EDUCATION

University at Buffalo, Buffalo — Aug. 2022 – Dec. 2023
Master of Science in Computer Science and Engineering — New York, US
Coursework: NLP, Computational Linguistics, Pattern Recognition, Machine Learning, Algorithms Design — GPA: 3.82 / 4

Indian Institute of Technology (Banaras Hindu University), Varanasi — Jul. 2014 – May 2018
Bachelor of Technology in Electronics Engineering — Uttar Pradesh, India

EXPERIENCE

Hilabs — Feb. 2024 – Present
Senior Data Scientist | Applied Data Science Team — Bethesda, MD

Legal Document Processing with GenAI

- Led the development of an advanced RAG-based Generative AI service aimed at processing 1000s of large legal documents efficiently for entity retrieval and verification of party agreements
- Leveraged both open-source LLMs (llama, mistal-ai) and enterprise-grade LLMs (GPT-4, Claude) to construct the solution while maintaining a paramount focus on data privacy
- Orchestrated the deployment of a GPU pipeline on Amazon EMR, ensuring inference times of less than 5 seconds using Docker/Terraform. Conducted rigorous load testing with 20 concurrent users to evaluate performance accurately
- Successfully completed a Proof of Concept (POC) which demonstrated the solution's effectiveness and potential for widespread adoption. Currently, integrating the solution for 3 clients, showcasing its impact and scalability

Apexanalytix — May. 2023 – Feb. 2024
Data Scientist | ML Engineer Intern | Analytics Team — Remote, New York

Generative Knowledge-Specific Chatbot

- Developed advanced Retrieval Augmented Generation(RAG) Chatbot with LLM for intelligent knowledge access across teams
- Engineered custom document chunking and indexing pipeline using LangChain and leveraged MiniLM embeddings
- Designed efficient RAG pipeline with ChromaDB, integrating MMR scoring and Azure OpenAI LLM to deliver precise response
- Achieved 89% approval in human evaluations and integrated technology into 12 internal and 18 external client applications

Oracle — Sept. 2020 – Aug. 2022
Senior Application Engineer(ML) | Oracle Financial Machine Learning Team — Bengaluru, India

Preemptive Anomaly Prediction in Corporate Billing

- Implemented in-memory multivariate anomaly prediction system for corporate billing, addressing monthly billing challenges
- Leveraged Oracle in-database ML for Semi-Supervised classification with both local and global model explainability
- Optimized service with indexing and parallelism, processing 1.2M bills and 5M segments in 20 mins with 92% precision
- Integrated system with services of Oracle Revenue Management and Billing (ORMB), granted innovation patent at USPTO

Wipro Limited — June 2018 – Sept 2020
Project Engineer(AI) | AI Research Team — Bengaluru, India

Chatbot Services for Employee Helpline Portal's Ticketing System

- Developed Employee Helpline Portal Chatbot with effective retrieval of historical ticket resolutions to enhance user support
- Improved user experience with precise responses using query intent classification and BERT-powered semantic search
- Achieved 79% accuracy score, reduced user wait times from 18 to 4 minutes, and decreased the workload for support agents

PATENTS and PUBLICATIONS

Paper Published: Virtual Conversation with Real-Time Prediction of Body Moments/Gestures — **ICMLIP 2019**
*Gopichand Agnihotram, Rajesh Kumar, Pandurang Naik, **Rahul Yadav*** — springerLink

Patent Granted: Method And System For Multimodal Analysis Based Emotion Recognition — **16/795840** [PDF]
*Inventors: **Rahul Yadav**, Gopichand Agnihotram* — Applicant: **Wipro Limited**

SKILLS

Languages : Python, Java, C++, C, PLSQL, SQL, MongoDB

Technologies : Retrieval Augmented Generation(RAG), Large Language Models(LLMs), Natural Language Processing(NLP), Machine Learning, Deep Learning, Web Services, Data Structures, Algorithms, Prompt Engineering, Indexing, Quantization

Cloud : Amazon Web Services (AWS), Azure, Google Cloud Platform (GCP), Oracle, Atlas, OpenAI

Frameworks and Libraries : Pytorch, Tensorflow, MLFlow, Kafka, SentenceTransformers(embeddings, re-rankers), Databricks, Snowflake, PySpark, Langchain, LlamaIndex, VectorDBs(ChromaDB, Faiss, elasticsearch), MLFlow, Git, Jenkins, Big Data, Hadoop, Flask, Docker, Kubernetes, Pandas, Streamlit, Flask-RESTful, FastAPI, YOLO, Django, XGBoost, GAN, ActiveMQ, Springboot