Praveen Kumar Varkala

pvarkala33@gmail.com | +1 (980) 307-4046 | Charlotte, NC | LinkedIn | GitHub | Portfolio

EXPERIENCE

United Health Group — **Jul 2024 – Present**
AI Engineer — Remote, USA

- Developed a claims explanation pipeline by fine-tuning LLMs such as GPT, BERT, LLaMA with LangChain and AI/ML technologies, achieving 95% accuracy in simplifying billing codes and reducing customer queries by 40%.
- Designed and deployed an AI-driven claims summarization tool using OpenAI APIs, Pinecone for vector retrieval, and AIOps, streamlining claim audits and improving operational efficiency.
- Implemented LoRA and QLoRA for parameter-efficient tuning and deployed AI models on AWS SageMaker, Kubernetes, and MLflow, ensuring scalability and cost optimization using AWS.
- Applied Computer Vision techniques to automate image analysis of claims documents, improving processing speed by 95%, integrating a Human-in-the-Loop (HITL) mechanism for quality control.
- Leveraged Looker and Google Analytics to monitor key performance indicators (KPIs) and identify areas for improvement in claims processing efficiency, resulting in XX% reduction in processing time.

Hexagon Capability Center India — **Aug 2022 – Dec 2022**
Software Engineer — Hyderabad, India

- Optimized Oracle SQL and NoSQL databases processing 10TB+ data, integrating Docker for scalable workflows.
- Built ETL pipelines in Snowflake, Redshift, and HDFS, processing 500GB+ daily, with AWS Lambda for automation.
- Applied regression models for SLA breach predictions with Datadog, improving on time delivery rates by 15%.

Citi Bank — **May 2019 – May 2022**
Software Engineer — Hyderabad, India

- Developed a full-stack web application using React and Typescript, implementing microservices and real-time data processing with Apache Kafka, improving response times by 30%.
- Designed RESTful and GraphQL APIs with JWT authentication, role-based access control, and OAuth 2.0, using Node.js, ensuring robust security, and API gateway protection in compliance with Citi's standards.
- Developed and integrated an NLP powered chatbot using RASA and Machine learning, automating customer inquiries, loan eligibility checks, and FAQs, achieving 95% intent recognition accuracy.
- Deployed and managed containerized applications using Docker and Azure Kubernetes Service (AKS), leveraging Terraform and Azure Functions for serverless, fault-tolerant, scalable, and cost-efficient solutions.
- Automated CI/CD pipelines using Azure DevOps and Jenkins, implementing secure file handling, enabling seamless code integration, and performing thorough Unit Testing and Regression Testing, reducing deployment time by 50%.

EDUCATION

University of North Carolina at Charlotte — **May 2024**
Master of Science in Computer Science; Concentration in Data Science — Charlotte, NC

ACADEMIC PROJECTS

RepoWise: AI-Powered GitHub Collaboration Platform

- Developed a full-stack SaaS platform using Next.js integrating Gen AI tools like Gemini and Assembly AI leveraging Foundational LLMs to summarize commit diffs, meeting recordings, and discussions, boosting team productivity.

End to End LLM Training and Optimization

- Built data pipeline for language model training, trained on 110M tokens with custom tokenizer and DeepSpeed and enhanced performance through Supervised Fine-Tuning (SFT), DPO, and RLHF with PPO.

OTHER

- **Languages and Frameworks**: Python, Java, Kotlin, React, Spring Boot, Hibernate, Remix, Next.js, Node, SQL.
- **Libraries**: TensorFlow, PyTorch, Keras, Streamlit, Langchain, Pandas, spaCy, NumPy, Transformers, XGBoost
- **Technologies**: Azure DevOps, RPA, Maven, AWS, Azure, GCP, MLOps, CI/CD, Gitlab, Airflow, Linux.