Praveen Kumar Varkala

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

EXPERIENCE

Probe Practice Solutions LLC — **Jul 2024 – Present**
AI Engineer — Round Rock, Texas

- Developed and evaluated a claims explanation pipeline by finetuning llama LLMs and Langchain, achieving 95% accuracy in simplifying billing codes into patient-friendly formats, reducing customer queries by 40%.
- Implemented and deployed a claims summarization tool using OpenAI APIs, Pinecone for vector retrieval, LLM Ops for best practices, maintaining 92% accuracy rate in summarization with CrewAI integration.
- Collaborated with product teams to develop and evaluate LLM based conversational AI solutions using Prompting techniques and Automatic Speech Recognition, achieving a 20% increase in IVR resolution rates.

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

- Optimized Oracle SQL databases processing 10TB+ data with Hive, integrating Docker for scalable workflows.
- Built ETL pipelines in Snowflake 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 Spring Boot, React and GraphQL, implementing microservices and real-time data processing with Apache Kafka, improving response times by 30% enabling live transaction updates.
- Designed GraphQL and RESTful APIs with JWT authentication, role-based access control (RBAC), and OAuth 2.0, ensuring robust security, encrypted storage, and API gateway protection in compliance with Citi's standards.
- Developed and integrated an AI-powered chatbot using RASA, automating customer inquiries, loan eligibility checks, and FAQs, achieving 95% intent recognition accuracy and reducing support workload by 40%.
- Deployed and managed containerized applications using Docker, Kubernetes, and Terraform, ensuring fault-tolerant, scalable, and cost-efficient cloud infrastructure on AWS.
- Automated CI/CD pipelines using Jenkins and GitHub Actions, enabling seamless code integration, automated testing, and efficient deployments, reducing deployment time by 50% and improving development agility.
- Contributed to system design and architecture by implementing scalable microservices, design patterns, and reliability improvements, ensuring high availability and efficient scaling of enterprise applications.

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 and react integrating AI tools like Google Gemini and Assembly AI 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, TypeScript, Java, C++, React, Flask, Next.js, Node, SQL, PostgreSQL
- **Libraries**: TensorFlow, PyTorch, Keras, Streamlit, Langchain, Pandas, spaCy, NumPy, Transformers, XGBoost
- **Technologies**: Machine Learning, Feature Engineering, AWS, MLOps, Predictive Modeling, Statistical Analysis, Supervised Fine Tuning, PEFT, Vector databases, RAG, LLM architecture design, LLM Model Evaluation.