Praveen K. Varkala

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

EDUCATION

University of North Carolina at Charlotte **Jan 2023 - May 2024**
Master of Science in Computer Science; Concentration in Data Science Charlotte, NC

EXPERIENCE

Meta **Jun 2025 – Present**
Software Engineer II (Imagine ML Team) **Remote, USA**

- Led evaluation and productionization of a multi-task diffusion model enabling precise image edits via natural language instructions, enhancing user engagement and retention by 40%
- Built a web-based evaluation interface in Next.js, allowing research teams to run inference, annotate outputs, and submit feedback
- Engineered Hive storage pipelines for 10M+ inference results, enabling efficient retrieval, and model comparison across evaluation runs
- Collaborated across research and infra teams on multiple image and video generation models, contributing features for multimodal evaluation and scaling to 100K+ generations/day

Global Atlantic Financial Group **Apr 2025 – Jun 2025**
Senior Software Engineer **Boston, MA**

- Deployed a RAG platform using Next.js, FastAPI, AWS Bedrock, and PostgreSQL, providing role-based access to internal databases and reducing employee query resolution time by 40%
- Implemented a text-to-SQL engine with LLMs fine-tuned for enterprise schemas, automatically generating SQL queries executed against internal data warehouses and feeding results into BI workflows
- Launched a dashboard builder module using Chart.js where users could combine multiple charts, edit, share, and export visualizations

Probe Practice Solutions **Jul 2024 – Mar 2025**
AI Engineer **Remote, USA**

- Launched a claims explanation pipeline by fine-tuning GPT LLMs, with LangChain and RAG pipelines, achieving ~65% accuracy in simplifying billing codes and reducing customer queries by 40%
- Designed AI-driven claims summarization tool using OpenAI APIs, Pinecone for vector retrieval, reducing manual claim audits by 60%
- Ensured compliance with HIPAA by integrating bias monitoring, model interpretability, and regulatory adherence using DeepChecks, increasing fraud detection accuracy by 50% and enhancing claims validation
- Collaborated with product and UX teams to design and implement user-friendly features, leading to a 40% increase in user engagement and a 30% improvement in customer satisfaction ratings
- Applied design patterns and participated in code reviews frequently to maintain high code quality, ensuring 95% adherence to best practices and reducing production issues by 35%

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

- Improved performance of a C# 3D visualization module by reducing model load times by ~30%, achieved through memory optimization and multithreading techniques
- Reduced defect backlog by ~20% by refactoring legacy code into reusable, modular components and object-oriented design patterns
- Increased system scalability by implementing efficient data-handling layers that integrated plant metadata with SQL databases, enabling faster retrieval and visualization of thousands of records

Citi Bank **Dec 2020 – May 2022**
Software Engineer **Hyderabad, India**

- Architected a secure, cloud-native banking platform using React, Spring boot, and hibernate integrated with Apache Kafka for event-driven architecture, improving system scalability and reducing response times by 30%
- Integrated an NLP-powered chatbot using RASA and spaCy, automating loan eligibility check, achieving 95% intent recognition accuracy
- Optimized database queries and caching with PostgreSQL, Redis and automated CI/CD pipelines using Jenkins and Terraform, reducing deployment time by 50%

SKILLS

- **Languages and Databases:** Python, Java, C++, JavaScript, TypeScript, MySQL, PostgreSQL, MongoDB, Kafka, Redis
- **Web technologies**: React, Next.js, Tailwind CSS, Spring Boot, Hibernate, Node.js, REST, GraphQL, JWT, gRPC, OAuth
- **ML Libraries and Technologies**: TensorFlow, PyTorch, Keras, Langchain, Pandas, spaCy, NumPy, Transformers, Feature Engineering, Predictive Modeling, Statistical Analysis, LLMs, SFT, PEFT, Vector databases, RAG
- **Cloud and DevOps**: AWS (S3, Lambda, Bedrock), Jenkins, Gitlab, Docker, Kubernetes, Terraform

ACADEMIC PROJECTS

RepoWise: AI-Powered GitHub Collaboration Platform

- Innovated 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

- Implemented a data pipeline to process datasets for language model training, created a custom tokenizer, and trained a model using DeepSpeed on 8 A100 GPUs, handling 110 million tokens