Praveen V.

reachpraveenvk@gmail.com | (650) 623-2823 | Jersey City | 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 **Remote, USA**

- Architected diffusion model for personalized multimodal image and video generation and instruction-based extensions, improving CLIP similarity by 7%, reducing FID by 29%, cutting latency by 60% through quantization
- Extended internal DAG orchestration system with multimodal nodes & automated workflow triggers, reducing manual setup time by 40%
- Engineered inference validation pipeline with human-in-the-loop annotation system for multimodal outputs, allowing 5+ research teams to label and submit feedback on 100K+ daily generations
- Implemented RLHF using direct preference optimization with MLflow and Weights & Biases, reducing model alignment iterations by 25% and deployed drift monitoring across 100K+ daily generations

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

- Designed RAG system with LangChain and FAISS vector store, retrieving relevant SQL examples from 50K queries, improving GPT-4 text-to-SQL accuracy from 22% to 62%
- Developed Next.js and FastAPI analytics platform with role-based access, automated SQL execution, and Chart.js charts, cutting management query turnaround from 2-3 days to minutes
- Launched AWS Bedrock scheduled jobs for automated chart data refresh across 50+ saved dashboards, maintaining visualization accuracy as enterprise data updates

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

- Fine-tuned GPT-3.5-turbo using LoRA (rank=8, alpha=16) on 50K medical claims, achieving 65% exact-match accuracy and 0.82 F1 in billing code simplification, reducing support tickets by 40%
- Integrated RAG pipeline with Pinecone and LangChain, retrieving ICD-10 and CPT medical ontologies with cross-encoder reranking, improving retrieval precision from 0.54 to 0.71
- Released fraud-detection model with SHAP explainability ensuring HIPAA compliance, improving F1 from 0.68 to 0.83 and reducing false positives by 30%

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

- Optimized spatial data queries using KD-tree indexing and batch processing, reducing retrieval time by 30% for 3D plant visualization with 100K+ geometric entities
- Refactored ETL pipelines for sensor data with PySpark, implementing feature extraction that accelerated analytics processing by 25%

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

- Built conversational AI system using RASA framework with DIET architecture for loan eligibility inquiries and customer FAQs, achieving 60% intent accuracy
- Deployed chatbot on AWS Lambda with Docker and PostgreSQL backend, processing 10K+ queries monthly across banking channels
- Created ML-based fraud detection system with Flask and PostgreSQL, improving detection accuracy by 12% and reducing false positives by 25%, enhancing transaction security for 100K+ accounts

SKILLS

- **Languages, Databases, & Messaging:** Python, Java, C#, JavaScript, TypeScript, PostgreSQL, MySQL, MongoDB, Redis, Kafka
- **Frameworks & Libraries**: React, Next.js, Spring Boot, FastAPI, TensorFlow, PyTorch, Keras, LangChain
- **Machine Learning**: Transformers, Retrieval Augmented Generation, feature engineering, statistical analysis, Finetuning (SFT, PEFT)
- **Cloud & DevOps**: AWS (S3, Lambda, Bedrock), Docker, Kubernetes, Jenkins, GitLab, Terraform

ACADEMIC PROJECTS

ArxivAI: Hybrid Search Research Platform | Link

- Architected multi-agent RAG system with OpenSearch hybrid retrieval (BM25 + Jina embeddings), section-based chunking, and Llama 3.2 LLM, achieving 78% accuracy in research paper summarization across 10K+ documents
- Implemented production optimizations with Langfuse monitoring and Redis caching, achieving 150x performance improvement for repeated queries and reducing average response time from 120s to 15s

RepoWise: AI-Powered GitHub Collaboration Platform | Link

- Built a full-stack SaaS platform using Next.js integrating AI tools like Google Gemini and Assembly AI to summarize commit diffs, meeting recordings, and discussions, boosting team productivity.
- Integrated functionalities using Octokit for GitHub API, enabling users to query the codebase contextually with AI-powered Q&A.