Praveen V.

650-623-2823 | reachpraveenvk@gmail.com | linkedin.com/in/praveenvarkala | github.com/PraveenKumarVk

Education

University of North Carolina at Charlotte — Charlotte, NC
Masters in Computer Science — *Jan 2023 – May 2024*

Experience

Software Engineer II — June 2025 – Present
Meta — *Remote, US*

- Built evaluation pipeline generating 100K+ daily inference outputs for 5+ multimodal diffusion models
- Automated batch generation and hyperparameter sweeps, reducing tuning cycle time by 25%
- Extended orchestration DAG with multimodal node support, increasing experiment throughput by 40%
- Delivered 7% CLIP improvement, 29% FID reduction by optimizing data quality and model parameters

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

- Implemented RAG with LangChain and FAISS over 50K queries, improving text-to-SQL accuracy by 40%
- Automated domain selection by building classification layer, routing user questions to relevant schema context
- Developed Next.js and FastAPI platform with role-based access for SQL execution and Chart.js visualization
- Deployed AWS Bedrock jobs to refresh dashboard data daily, updating over 50+ charts

AI Engineer — July 2024 – Mar 2025
Probe Practice Solutions — *Remote, US*

- Fine-tuned GPT-3.5 using LoRA to simplify billing codes for patients, reducing support tickets by 40%
- Implemented RAG with Pinecone and LangChain for CPT medical codes, improving lookup accuracy by 30%
- Improved fraud detection model F1 score by 25% using SHAP for explainable predictions in HIPAA compliance

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

- Reduced 3D plant visualization load time by 30% by implementing KD-tree indexing and batch processing
- Implemented feature extraction workflows and ETL pipelines with PySpark, improving data processing by 25%

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

- Built conversational AI chatbot using RASA, handling 5K+ loan eligibility inquiries and customer FAQs daily
- Deployed chatbot on AWS Lambda processing 50K+ monthly queries, achieving 60% intent recognition accuracy
- Improved fraud detection system with continuous monitoring and analyst feedback, reducing false positives by 25%

Projects

ArxivAI | *OpenSearch, Langfuse, FastAPI, Docker, Redis* — Sep 2025 – Present

- Architected multi-agent RAG system with OpenSearch hybrid retrieval, section-based chunking, and Llama 3, achieving 78% accuracy in research paper summarization
- Optimised using Langfuse monitoring and Redis caching, achieving 150x performance improvement for repeated queries and reducing average response time to 15s

RepoWise | *TypeScript, LangChain, Google Gemini, Assembly AI, Octokit* — Oct 2024 – Dec 2024

- Developed 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

Technical Skills

Languages, Databases, & Messaging: Python, Java, TypeScript, PostgreSQL, MySQL, MongoDB, Redis, Kafka
Frameworks & Libraries: React, Next.js, FastAPI, PyTorch, LangChain, HuggingFace, Transformers
Machine Learning: transformers, RAG, feature engineering, prompt engineering, fine-tuning (SFT, PEFT)
Cloud DevOps: AWS (S3, Lambda, Bedrock), Docker, Kubernetes, Jenkins, GitLab, Terraform