Abhinaav Ramesh

San Jose, CA | contactabhinaav@gmail.com | (213)-421-9945 | Linked-In

Resourceful AI/ML engineer passionate about transforming advanced research into production systems. Energetic problem-solver who thrives in high-ambiguity, cross-functional environments. Expert in architecting complex agentic AI systems, optimizing LLM training and inference workflows, and scaling distributed architectures that bridge foundational models with real-world impact.

EXPERIENCE

Hewlett Packard Enterprise — San Jose, CA
Data Scientist II - Agentic AI & LLM Systems — January 2023 – Present

- Architected and developed Networking Copilot with Agentic Mesh: Production-scale hierarchical multi-agent system using LangGraph serving 6M+ devices, 3B+ clients, 100K+ customers globally with 50% MTTR reduction and 88% CSAT; projected to drive 20x licensing revenue increase
- Shipped production multi-agent orchestration at scale with ReAct reasoning patterns, supervisor-worker coordination, and multi-tier memory systems deployed across distributed, multi-tenant real-time streaming infrastructure handling concurrent conversational AI interactions with load balancing, rate limiting, and distributed model inference
- Engineered custom Multi-Task BERT architecture for Search, jointly optimizing ranking, intent/entity tagging, and snippet selection with 98.5% accuracy, achieving lower latency and higher relevance for enterprise network search (HPE Invention, Patent Pending)
- Designed Agent evaluation LLM-as-judge frameworks with scoring for reasoning quality, task completion and accuracy assessment
- Built synthetic data generation pipelines for fine-tuning and pre-training SLMs and LLMs on domain-specific tasks optimizing cost and latency
- Implemented embedding-based semantic anomaly detection for network traffic analysis, transforming >2B application flow records (ASN-enriched traffic data) into semantically dense vector spaces using custom NLP-based MapReduce functions, enabling real-time threat detection and behavioral clustering at scale
- Deployed observability pipelines on MLFlow, LangFuse and OpenTelemetry for distributed tracing and explainability; generated labeled datasets for reinforcement learning from human feedback (RLHF)

USC Mark and Mary Stevens Neuroimaging and Informatics Institute — Los Angeles, CA
Graduate Researcher - Deep Learning and Analytics — August 2021 – April 2023

- Developed encoder-decoder architectures and StyleGAN-based generative models for 3D brain MRI segmentation, achieving >95% IOU and building production pipelines with custom data loaders and loss function
- Built Auto-QC framework for segmentation quality and machine learning model performance for shape2vec
- Built image processing libraries to model 3D brain morphology trajectories, creating custom libraries for advanced 3D image processing.
- Co-authored 6 peer-reviewed publications on data modelling and deep learning for medical imaging

Bosch Global Software Technologies — Bengaluru, India
Member Technical Services - AI — December 2019 – June 2021

- Shipped AI at edge for digital pathology: Developed GAN-based image enhancement model for whole slide imaging microscopy achieving >90% accuracy, optimizing embedded system algorithms for 40% reduction in focus time on z-axis precision measurement
- Deployed ML algorithms to AWS Cloud: Optimized deep learning models for cloud inference, increasing execution speed by 35% in production deployment; collaborated with end-users (pathologists) to drive hardware/software design improvements

TECHNICAL SKILLS

- **Agentic AI**: LangGraph, LangChain, MCP, A2A, Multi-Agent Orchestration, RAG, RLHF, LLM-Ops, LLM-as-Judge, Prompt Engineering
- **LLM Fine-tuning & Optimization:** PEFT (LoRA, QLoRA), Prefix Tuning, Instruction Tuning, Quantization, Model Distillation
- **ML/DL Frameworks**: PyTorch, TensorFlow, JAX, Scikit-Learn, XGBoost, Transformers, DeepSpeed, Accelerate, CUDA
- **Cloud & Infrastructure**: AWS Bedrock, OpenSearch Retrieval, NL2SQL, EMR, Docker, Kubernetes
- **Observability & Evaluation**: OpenTelemetry, MLFlow Tracking, LangFuse, Distributed Tracing, Custom Eval Pipelines
- **Data Science -** Statistical Analysis, Feature Engineering, Data Modeling, Stochastic Models, A/B Testing

KEY PUBLICATIONS & ACHIEVEMENTS

- HPE Discover 2025 Keynote: Led development of Aruba Networking Copilot with Agentic Mesh, featured in CEO Antonio Neri's keynote as industry-first autonomous network operations solution – Agentic Mesh, trained on trillions of data points from millions of devices
- Peer-reviewed papers in IEEE, EMBC, NER, SIPAIM, MLMI, ICMLAN, Nature on AI (ML / DL) and computer vision *Google Scholar*
- Hackathon Winner: Cal Hacks, PennApps, GRIDS USC – Data Analytics for health, AI at edge devices , Transformers for ASL
- Patent Pending: Deep learning algorithm for mechanical deviation compensation in digital pathology (Bosch)
- Certifications: Microsoft Professional Program in Artificial Intelligence, Google Data Analytics Professional Certification, DeepLearning.ai (Deep Learning, GANs, LangChain, AI Agents in LangGraph, Functions/Tools/Agents, Finetuning LLMs)

EDUCATION

University of Southern California, Los Angeles, CA | Master of Science in Data Science — *May 2023*
Anna university, India | Bachelor of Engineering in Biomedical Engineering — *September 2020*