Chinmay Appa Rane

Austin, Texas | (682) 583-1880 |drranechinmay@gmail.com | https://chinmayrane.com/ |Linkedin | GitHub

SUMMARY

Ph.D. ML Engineer with 7+ years delivering enterprise AI impact—$50K+/day savings, 40% accuracy improvements—across LLMs, VLMs, and computer vision. ICLR published, NVIDIA GTC speaker.

TECHNICAL SKILLS

Languages	Python, C++, MATLAB, SQL
ML/AI Frameworks	PyTorch, TensorFlow, Keras, MONAI, Hugging Face Transformers, Scikit-learn, DeepSpeed, LLaMA Factory, Unsloth, LangChain, CrewAI, LangGraph (learning)
AI/ML Techniques	Computer Vision (Segmentation, Object Detection), LLMs, Multimodal VLMs, RAG, Self-Supervised Learning, Federated Learning, Transformers, Diffusion Models, PEFT, Full Supervised Fine Tunning
Production & Deployment	TensorRT, ONNX, ONNX Graph Surgeon, vLLM, TensorRT-LLM, Triton Inference Server, Docker, Docker Compose, Kubernetes, FastAPI, REST APIs
Cloud & MLOps	AWS (SageMaker, EC2, S3), GCP, CI/CD, MLflow, Model Versioning, Milvus Vector DB, Prometheus, Grafana
Domains	Medical Imaging, Autonomous Systems, Generative AI, Technical Leadership

WORK EXPERIENCE

Senior Machine Learning Engineer - *Quantiphi Inc* *(September 2021- Current)*

Leading AI development across Generative AI, multimodal AI/ML (LLMs, VLMs, computer vision, medical imaging) from research to production, delivering scalable solutions with hands-on technical leadership.

- **Structural Repair Document Creation using VLMs & LLMs** – Built VLM/LLM system (InternVLM, LLaMA 3.2, RAG) reducing helicopter ground time by 70% ($50K+ daily savings) through automated damage detection and repair report generation. Led technical implementation and customer relations.
- **Clinical LLM Model Building and Optimization –** Engineered DeepSpeed/Unsloth training pipeline reducing LLaMA fine-tuning time by 55% (42→19 hours, $15K+ savings/cycle) across 8x H100 GPUs (95%+ efficiency). Implemented vLLM inference for production scale while mentoring team on distributed training.
- **3D Texture Generation using Vision-Language** – Developed VLM-based automation reducing texturing time by 80% (40→8 hours) using CLIP embeddings with DBSCAN clustering for zero-shot component recognition. Co-led dual-deployment architecture (laptop/server) for manufacturing workflows.
- **Enterprise-grade AI platform for cardiovascular diagnostics** – Deployed self-supervised MONAI pipeline reducing annotation costs 10× ($50K→$5K) and report generation time by 50% with 40% accuracy improvement. Production system on AWS/Triton serving 7 hospitals at 85% accuracy for surgical planning.
- **RCM AI Agent for Claims Denial Resolution(Ongoing) –** Developed AI agent system (LangChain, vLLM) automating healthcare claims denial resolution, reducing manual processing time by 30-40% through intelligent workflow automation. Led POC implementation with custom tool creation for CARC/RARC code verification.

ADDITIONAL WORK EXPERIENCE

Neural Network Research Assistant - *Image Processing and Neural Networks Lab, UTA* *(2017- 2021)*

Developed custom neural network algorithms for medical imaging (LASIK surgery) and geophysical applications using C++, MATLAB, and Python.

Data Scientist Intern - *Unique Software Development* *(Jan 2018 – Dec 2018)*

Built production NLP pipelines (Amazon Comprehend) and object detection systems (YOLO, TensorFlow Lite) for transportation analytics and robotics applications.

ADDITIONAL PROJECTS

Personal projects include RAG semantic search, LLM agent workflows (CrewAI, LangChain), real-time object detection with automated reporting, and edge AI deployment (Raspberry Pi home automation).

**For detailed project descriptions and additional work, visit https://chinmayrane.com/.*

EDUCATION

University of Texas at Arlington | Doctor of Philosophy, *Deep Neural Networks* | Electrical Engineering *2021*

University of Texas at Arlington | Master of Science, *Neural Networks* | Electrical Engineering *2016*

TALKS AND RECOGNITION

NVIDIA GTC 2025 Speaker *– Presented "Revolutionizing Cardiac MRI Analysis and Diagnosis With AI: A Deep Dive into MONAI-Based Cardiac MRI Segmentation" under NVIDIA's Healthcare AI track.*

PUBLICATIONS

- Chinmay Rane, Michael Manry, "Dynamic Activations for Neural Net Training", The Second Tiny Papers Track at ICLR 2024 .
- Chinmay Rane, Sanjeev Mallur, Yash Shinge, Kanishka Tyagi, Michael Manry, "*Optimal Input Gain: All You Need to Supercharge a Feed-Forward Neural Network*", *ArXiv.*
- Kanishka Tyagi, Xun, Chinmay Rane, Michael Manry, "Automated Sizing and Training of Efficient Deep Autoencoders using Second Order Algorithms", *ArXiv.*