Chinmay Appa Rane

Austin, Texas | (682) 583-1880 | drchinmayrane@outlook.com | https://chinmayrane.com/ |Linkedin

SUMMARY

Ph.D.-level ML Engineer combining 7+ years of AI research and production expertise in Generative AI, LLMs, and Computer Vision. Technical leader delivering measurable enterprise impact - $50K+/day savings, 40% diagnostic accuracy improvements - with publications in ICLR and presentations at NVIDIA GTC.

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

Production & Deployment: TensorRT, ONNX, ONNX Graph Surgeon, vLLMs, 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 and deployment by combining expertise in Generative AI, multimodal AI/ML (LLMs, vision-LLMs, computer vision, and medical imaging) with strategic leadership and project management. Delivering scalable AI solutions across industries, from research to production.

- ○ **Clinical LLM Model Building and Optimization –** Architected enterprise-grade LLM fine-tuning and inference pipeline using DeepSpeed, Unsloth, and LLaMA-Factory, reducing training time by 55% and saving $15K+ per run. Deployed optimized inference with vLLM across 8×H100s, maximizing LLaMA-4 memory utilization.
- ○ **3D Texture Generation using Vision-Language** – Architected VLM-driven 3D texture automation system for aerospace, cutting designer effort by 80% per aircraft component. Combined CLIP embeddings with DBSCAN clustering for zero-shot component recognition. Deployed flexible dual-mode solution (laptop + server) via Docker Compose.
- ○ **Structural Repair Document Creation using VLMs & LLMs** – Led development of a visual damage detection and repair documentation system using InternVLM and RAG, reducing aircraft ground time by 70% and saving $50K+/day. Integrated repair manuals for automated recommendations, significantly cutting manual tasks. Managed stakeholder alignment and program delivery.
- ○ **Enterprise-grade AI platform for cardiovascular diagnostics -** Led development of a MONAI-based self-supervised segmentation pipeline, reducing cardiologist report time by >50% and improving diagnostic accuracy by 40%. Designed architecture to cut annotation costs 10× (from $50K to $5K) while maintaining > 85% accuracy. Deployed on AWS with Triton Inference Server across 7 hospitals, enabling real-time surgical planning integration.

ADDITIONAL PROJECTS

- **Semantic Search RAG –** Python, Docker, Ollama, Hugging face, Retriever, Re-ranker, RAG.
- **Weather Data Summarization using LLM Agents –** Python, CrewAI, Ollama, Docker, Hugging face, LLM Agents, Multimodal LLMs, LLaVA.
- **Real-time Object Detection & Automated Incident Reporting with LLM Agents –** Python, YOLOv8, Flask, FastAPI, CrewAI, Ollama, LLaVA, Docker, RESTful APIs, Base64 Encoding, LLM Agents, Multimodal LLMs.
- **Home Assistant LLM Voice Agent (Edge Deployment Project)** *Personal Project | Raspberry Pi + Home Server Integration -* Python, Ollama, LLaMA 3.2 3B, Whisper, Piper, Home Assistant, Wyoming Protocol, Docker, Raspberry Pi.

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

EDUCATION

University of Texas at Arlington | Doctor of Philosophy, *Deep Neural Networks* | Dept – EE *(Spring 2021)*

University of Texas at Arlington | Master of Science, *Neural Networks* | Dept – EE *(Spring 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.*

ADDITIONAL WORK EXPERIENCE

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

Conducted research and developed custom neural network algorithms for image processing, signal processing, and AI applications. Contributed to industry-funded projects involving AI for LASIK surgery and geophysical imaging.Developed prototypes in C++, MATLAB, and Python integrating classical and deep learning methods.

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

Developed Python-based production grade code for NLP using Amazon Comprehend, custom ML models, and object detection algorithms to process complex user data. Integrated advanced querying, analytics, and predictive tools for Mode Transportation. Additionally, implemented YOLO and face detection algorithms for the Inmoov 3D robot and optimized object detection on a Raspberry Pi using TensorFlow Lite.

PUBLICATIONS

- Chinmay Rane, Michael Manry, "Dynamic Activations for Neural Net Training", The Second Tiny Papers Track at ICLR 2024 .
- Kanishka Tyagi, Chinmay Rane, Bito Irie, Michael Manry, "*Multistage Newton's approach for training radial basis function neural network*", *SN Computer Science*, Publish date - June 2021.
- Kanishka Tyagi, Chinmay Rane, Michael Manry "*Regression analysis, Artificial Intelligence and Machine Learning for Edge Computing*" to be published by *Elsevier*, Accepted, Publish date - late 2021.
- Kanishka Tyagi, Chinmay Rane, Michael Manry "*Supervised Learning, Artificial Intelligence and Machine Learning for Edge Computing*" to be published by *Elsevier*, Accepted, Publish date - late 2021.
- Kanishka Tyagi, Chinmay Rane, Michael Manry "*Unsupervised Learning, Artificial Intelligence and Machine Learning for Edge Computing*" to be published by *Elsevier*, Accepted, Publish date - late 2021.
- Chinmay Rane, Sanjeev Mallur, Yash Shinge, Kanishka Tyagi, Michael Manry, "*Optimal Input Gain: All You Need to Supercharge a Feed-Forward Neural Network*", *ArXiv*, Publish date -April 2023.
- Kanishka Tyagi, Xun, Chinmay Rane, Michael Manry, "Automated Sizing and Training of Efficient Deep Autoencoders using Second Order Algorithms", arXiv preprint arXiv:2308.06221.