# Butchi Venkatesh Adari

butchivenkatesh.a@gmail.com +1(774)-670-7192 github.com/VenkateshRoshan linkedin.com/in/abven venkateshroshan.github.io

## Technical Skills

**Languages:** Python, C++, JavaScript, SQL, CUDA, PostgreSQL
**Frameworks & Libraries:** PyTorch, TensorFlow, Scikit-learn, Hugging Face, LangChain, LangGraph, FastAPI, ROS, PySpark
**Tools & Infra:** Docker, Kubernetes, MLflow, DVC, Git, Prometheus, Grafana, Ne04j, AWS, GCP, Azure
**ML Ops & Deployment:** Model Deployment, CI/CD, Inference Optimization, Vector Databases (FAISS)
**Course Knowledge:** Motion Planning, Computer Vision, DL, LLM, VLLM, MLOps, Foundations of Robotics, Robot Dynamics

## Experience

**Machine Learning Engineer Intern** *July 2025 - Present*
***Sellwizr***
- Developing and deploying LLM-powered Entity Resolution pipelines integrating PostgreSQL, OpenSearch, and Neo4j to achieve high-accuracy record matching across structured and unstructured datasets.
- Implementing blocking, and embedding-based retrieval to optimize candidate generation and improve pipeline efficiency.

**Graduate Researcher** *Worcester, Massachusetts*
***ELPIS LAB*** | *Worcester Polytechnic Institute* *Aug 2023 - May 2025*
- Improved monocular depth estimation for robotic grasping by reducing RMSE by 70% through distributed supervised fine-tuning and custom loss tuning, enabling reliable grasping in scenes where RealSense failed.
- Designed a Grasp Transformer for joint depth, pose, and heatmap prediction, achieving 65% grasp success in cluttered short-range scenes by combining monocular input with LangSAM-based segmentation.
- Achieved ±1–2cm depth accuracy for objects under 30cm and completed successful grasps using monocular predictions, where depth sensors returned no data.
- Deployed a 3 FPS real-time grasp inference pipeline using PyTorch and ROS2, integrating modular APIs, hand-eye calibration routines, tf broadcasters, and ROS drivers for consistent closed-loop execution in robotic systems.

**Machine Learning Engineer** *Hyderabad, India*
***Tata Consultancy Services*** *July 2021 - June 2023*
- Built a scalable OCR-NLP pipeline using transformer-based models TrOCR and LayoutLM with CRFs, processing over 600 scanned forms per hour to extract structured data and support downstream fraud analysis.
- Deployed the automated extraction pipeline on AWS using Lambda functions, achieving 94% structured data accuracy and enabling scalable cloud-based document processing.
- Developed a real-time people tracking system using YOLOv5 and DeepSORT with ONNX and TensorRT, running at 25 FPS across CCTV streams on dual NVIDIA GPUs in a retail mall.
- Generated zone-level foot-traffic heatmaps every 15 minutes and performed ROI-based dwell-time analysis to identify high-engagement areas, supporting layout optimization and promotion planning.

## Projects

**AI-powered Resume Match Agent | AI, LLMs, RAG** *June 2025 - July 2025*
- Orchestrated a 4-agent LLM pipeline (resume, JD parsing, matching, advising) using LangGraph and OpenAI/Ollama, achieving 5 to 6s end-to-end latency per analysis.
- Dockerized and deployed the full-stack system (FastAPI + Gradio UI), enabling seamless local or cloud execution; supports PDF/DOCX/TXT input and delivers real-time match scores, skill gap insights, and tailored resume suggestions.

**Agent based Web Data Extractor for RAG Systems | AI, LLMs, Web Scraping** *Feb 2025 - Mar 2025*
- Engineered an LLM-powered web agent using LangChain for RAG pipelines using the LLaMA model via Ollama API, enabling local, real-time document retrieval through multimodal filtering and navigation.
- Constructed a scalable, multi-threaded crawler with content-aware logic to enhance document selection and support high-accuracy downstream retrieval tasks in a local setup.

**Research Paper QA System with RAG Architecture and MLOps** *Oct 2024 - Dec 2024*
- Rolled out a RAG-based pipeline on GCP Vertex AI using CI/CD pipelines, ChromaDB for retrieval, and OpenAI's GPT-4 API for real-time scientific question answering.
- Achieved 1.3 second inference latency on CPU for interactive document-level QA with a scalable, production-ready pipeline.

**Image Captioning with Vision Transformer and GPT-2 | VLLM, NLP** *May 2024 - Jun 2024*
- Trained and fine-tuned a ViT,GPT-2 pipeline for image captioning using PyTorch, achieving 90% semantic relevance.
- Hosted the model on Hugging Face Spaces and streamlined deployment pipelines with GitHub Actions and AWS.

**High-Fidelity 3D Scene Reconstruction Using NeRF | Computer Vision** *Mar 2024 - Apr 2024*
- Reconstructed 3D scenes from 2D images, improving scene accuracy by 25% with and without positional encoding.

**Tesla Vision | Deep Learning, Computer Vision** *Jan 2024 - Feb 2024*
- Simulated a 3D autonomous driving dashboard using YOLO3D for vehicle detection, a custom lane recognition model, and ZoeDepth for monocular depth perception in an interactive driving scene in Blender with motion-prediction and visualization.

## Education

**Worcester Polytechnic Institute** *Aug 2023 - May 2025*
*Masters in Robotics Engineering - GPA: 3.8/4.0*

**Anil Neerukonda Institute of Technology and Sciences** *July 2017 - May 2021*
*Bachelors in Computer Science and Engineering - GPA: 7.77/10*