Butchi Venkatesh Adari

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

Education

Worcester Polytechnic Institute *Aug 2023 - May 2025*
M.S. in Robotics Engineering

Anil Neerukonda Institute of Technology and Sciences *Jul 2017 - May 2021*
B.Tech in Computer Science and Engineering

Technical Skills

Languages: Python, C++, JavaScript, SQL, Bash
Machine Learning: PyTorch, TensorFlow, Scikit-learn, Transformers, Pinecone, MCP, LangChain
Cloud & ML Platforms: Google Cloud Platform (Vertex AI, GCS, BigQuery), AWS, Kubernetes, Docker, GitHub Actions, Kafka
Databases: PostgreSQL, MySQL, MongoDB, NoSQL, Redis, Vector Databases
Backend Technologies: Django, Flask, FastAPI, RESTful APIs, GraphQL, Nginx
Libraries & Tools: Pandas, NumPy, Matplotlib, Git, GitHub, GitLab CI, Docker, REST API

Experience

Founding Machine Learning Engineer *NewYork, NY*
Alpheva AI *Aug 2025 - Present*

- Deployed a containerized multi-agent AI platform on AWS ECS with FastAPI microservices and async orchestration, supporting 2,000+ concurrent users with 99.9% uptime, optimizing end-to-end latency via async execution, streaming responses.
- Architected a stateless orchestration and routing layer using Redis Streams and SQS with retries, circuit breaking, and batching, reducing cascading failures by 70% and improving system stability during traffic spikes.
- Implemented a scalable data layer using PostgreSQL/SQL for user metadata, DynamoDB for conversation history, and Redis for low-latency caching, enabling fast personalization and reliable state management at scale.
- Built end-to-end ML pipelines covering data ingestion, retrieval, Vertex AI–based model inference, and RLHF-driven LLM fine-tuning pipelines (reward modeling, preference optimization), improving recommendation relevance by 20–25% and reducing incorrect outputs by 30%.
- Productionized full MLOps and observability with Prometheus, Grafana, MLflow, cutting incident detection time by 60% and enabling zero-downtime deployments across all micro-services.
- Increased LLM inference efficiency using Google Vertex AI with adaptive model routing across Gemini 2.5 Flash, 2.5 Pro, and 3 Pro, token-aware execution, batching, and RAG-based context selection, reducing inference costs by 40% while doubling throughput.

Graduate Researcher - Robotic Perception & Grasping *Worcester, MA*
ELPIS Lab, Worcester Polytechnic Institute *Aug 2023 - May 2025*

- Improved monocular depth estimation via multi-GPU VLM fine-tuning on the Grasp1B dataset, achieving 70% RMSE improvement over baseline and enabling successful grasps in scenarios where Intel RealSense depth sensors failed.
- Designed a Grasp Transformer predicting depth, pose, and grasp heatmaps directly from RGB, achieving 65% grasp success rate in cluttered, short-range manipulation tasks.
- Deployed FP16 (float16) inference for depth and grasp models on real robotic hardware, reducing latency and memory footprint to support fast, reliable monocular grasping from a single image.
- Built an end-to-end real-time perception → grasp pipeline (3 FPS) in PyTorch and ROS2 with hand–eye calibration and hardware drivers, ensuring stable execution under compute and runtime constraints.
- Validated robustness through simulation-to-real transfer using PyBullet synthetic data and physical trials, observing only a 10% performance drop and improved resilience to lighting and viewpoint variations.

Machine Learning Engineer *Hyderabad, India*
Tata Consultancy Services *Jul 2021 - Jun 2023*

- Engineered a scalable OCR–NLP pipeline using TrOCR, LayoutLM, and CRFs to extract structured fields from scanned documents, processing 600+ forms per hour for fraud and risk analytics.
- Productionized the document extraction system on AWS Lambda with S3-backed storage, achieving 94% structured-data accuracy and reducing manual data entry and review effort by 50%.
- Implemented a real-time people-tracking solution using YOLOv5 and DeepSORT, enhanced with ONNX and TensorRT, sustaining 25 FPS across CCTV streams to generate heatmaps and dwell-time metrics.
- Architected Python-based backend services to coordinate OCR, NLP, and vision inference workflows, increasing end-to-end pipeline throughput by 35% and enabling independent component scaling.
- Developed RESTful APIs and asynchronous job pipelines for document and video ingestion, inference management, and analytics delivery, cutting processing latency by 40% and simplifying downstream integrations.

Projects

EasyTex - AI agentic LaTeX Builder | Python, AWS, LangChain *June 2025 - Present*

- Constructed a production-grade multi-agent AI platform integrating multiple LLMs with load balancing and failover, supporting 500+ weekly active users while maintaining 99.9% availability and low-latency responses.
- Designed an AI-powered chat assistant using RAG with proprietary session memory and prompt optimization, delivering 30% faster resume analysis and 25% higher response accuracy across long multi-turn conversations

Tesla Vision | 3D Autonomous Driving Simulation | Computer Vision *Feb 2024 - Mar 2024*

- Developed a 3D autonomous driving simulation in Blender that fuses YOLO3D vehicle detection, custom lane-recognition, ZoeDepth monocular depth, and trajectory prediction into a cockpit-style dashboard for rapid perception prototyping.