Butchi Venkatesh Adari R

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

Education

Worcester Polytechnic Institute August 2023 – May 2025
M.S. - Robotics Engineering

Anil Neerukonda Institute of Technology and Sciences July 2017 – May 2021
B.E - Computer Science and Engineering

Experience

Founding Software Engineer | Python, React Native, RAG, MCP, AWS August 2025 – Present
Alpheva AI

- Deployed a containerized multi-agent AI platform on AWS ECS/EKS with FastAPI microservices and async orchestration, supporting 2,000+ concurrent users, achieving 99.9% uptime and stable 6–30s response times under peak load.
- Built 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.
- Integrated a production RLHF pipeline (reward modeling, PPO, rejection sampling) using live user feedback, improving recommendation relevance by 20–25% and reducing incorrect agent outputs by 30%.
- Productionized full MLOps and observability with Prometheus, Grafana, MLflow, and OpenTelemetry, cutting incident detection time by 60% and enabling zero-downtime deployments across all services.
- Increased LLM inference with adaptive routing, batching, and token-aware execution, reducing inference costs by 40% while doubling system throughput.
- Executed large-scale load and failure testing across microservices, validating system behavior beyond 2 × expected traffic and preventing downstream outages.

Software Development Engineer - II | Python, Computer Vision, AWS, ONNX July 2021 – June 2023
Tata Consultancy Services *Hyderabad, India*

- 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.
- Optimized model serving through batch inference and efficient serialization, lowering AWS compute costs by 30% while preserving real-time and near–real-time SLAs.
- Established monitoring and structured logging for production pipelines, improving failure detection and operational reliability for continuously running OCR and video analytics workloads.

Projects

EasyTex - AI agentic LaTeX Builder | Next.js, 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.
- Pioneered a multimodal autonomous interview agent combining TTS, STT, and an AI evaluation engine, enabling real-time rubric-based scoring and structured feedback across 100+ simulated interview sessions.

Robotic Monocular Grasping | Robotics, Grasp Transformers, ROS2, PyTorch August 2023 – May 2025

- Improved monocular depth estimation models for robotic grasping by 70% RMSE over baseline, enabling successful grasps in scenarios where Intel RealSense depth sensors completely failed.
- Formulated a Grasp Transformer architecture predicting depth, pose, and grasp heatmaps directly from RGB, achieving 65% grasp success rate in cluttered, short-range manipulation tasks.
- Assembled an end-to-end 3 FPS grasping pipeline in PyTorch and ROS2 with hand–eye calibration and ROS drivers, allowing research models to run reliably on real robotic manipulators.
- Validated the grasping system through simulation-to-real transfer using PyBullet-based synthetic data and real-world trials, achieving 10% performance drop between simulation and physical deployment and improving model robustness across lighting and viewpoint variations.

Skills

Programming Languages: Python, C++, JavaScript, C, Java, Dart, TypeScript, SQL, HTML5, CSS3

Machine Learning: PyTorch, TensorFlow, Scikit-learn, Transformers, RAG, Lang Chain, GraphRAG, Pinecone, MCP

Cloud: AWS (EC2, CloudFormation, API Gateway, CloudWatch, IAM), Microsoft Azure, GCP, Kubernetes, Kafka, GitHub Actions

Databases: Oracle, PostgreSQL, MySQL, MongoDB, NoSQL, Apache Cassandra, Redis

Backend Technologies: Django, Flask, FastAPI, Node.js, Spring, gRPC, RESTful APIs, GraphQL, Nginx, Elasticsearch, Cassandra

Libraries & Tools: pandas, NumPy, Matplotlib, Figma, JIRA, Git, GitHub, GitLab CI, Jenkins, Docker, REST API