VAMSHI KRISHNA

USA (Open to Relocation) | +1 4752679591 | vamshikrishna81199@gmail.com | LinkedIn| GitHub

SUMMARY

AI/ML Engineer with 4+ years of experience building scalable GenAI and machine learning solutions across NLP, computer vision, and automation domains. Skilled in LLM fine-tuning, RAG, and agentic workflows using tools like LangChain, CrewAI, and Vertex AI. Proven ability to reduce deployment time and improve model accuracy through cloud-native CI/CD pipelines on GCP and AWS.

EXPERIENCE

GenAI Engineer, Prozech (UiPath), USA — Jan 2025 – Present

- Designed and deployed agentic AI workflows using LangGraph and CrewAI, enabling autonomous decision-making and reducing manual intervention in business processes by 40%.
- Integrated Retrieval Augmented Generation (RAG) with Pinecone and Snowflake Cortex, improving LLM response accuracy by 35% and reducing hallucination rates in production systems.
- Fine-tuned Transformer models (GPT, LLaMA) using Hugging Face and Vertex AI, improving customer intent classification accuracy from 82% to 93% on real-time chatbots.
- Built CI/CD pipelines for ML lifecycle on GCP and AWS SageMaker, cutting deployment time by 50% and enabling scalability.
- Developed multi-agent orchestration using AutoGen and LangChain, increasing task success rate by 45% in complex document processing and business rule automation scenarios.
- Teamed with data engineers and product teams to deploy NLP support automation, cutting resolution time by 30% and boosting satisfaction 25%.

AI/ML Engineer, Anvizon (PwC), India — Mar 2021 – Aug 2023

- Developed and deployed a real-time object detection model using TensorFlow and Keras for EO satellite images, improving target identification accuracy by 38%.
- Engineered a cloud-native ML pipeline on AWS (SageMaker + Lambda) to process SAR imagery, reducing data processing latency by 55% and supporting defense-grade scalability.
- Led a cross-functional AI team to deliver a deep learning-based image classification system with 92% precision for RF signal pattern detection in aerospace telemetry.
- Optimized AI models in PyTorch for geospatial image enhancement and denoising, boosting feature extraction accuracy by 47%.
- Contributed to AI/ML strategy roadmap by evaluating and integrating computer vision techniques aligned with mission-critical space programs, accelerating delivery timelines by 20%.

SKILLS

- **Generative AI & LLMs**: LangChain, LangGraph, CrewAI, AutoGen, Retrieval-Augmented Generation (RAG), Pinecone, Snowflake Cortex, Hugging Face, OpenAI, LLaMA, GPT, n8n AI
- **Machine Learning & Deep Learning**: TensorFlow, PyTorch, Keras, Vertex AI, AWS SageMaker, Image Classification, Object Detection, Fine-Tuning, Transfer Learning, Signal Processing, Satellite & SAR Image Analysis
- **Cloud & MLOps**: AWS (Lambda, SageMaker, S3), Google Cloud Platform (Vertex AI, Cloud Functions), CI/CD for ML, Model Deployment, ML Pipelines, Containerization (Docker)
- **Natural Language Processing**: NLP-based Automation, Intent Classification, Transformer Models, LLM Hallucination Mitigation, Prompt Engineering
- **Data Engineering & Ops**: Data Preprocessing, Feature Engineering, Geospatial Data Processing, Big Data Pipelines, Real-Time Data Ingestion
- **Collaboration & Strategy**: Cross-Functional Team Leadership, Product Integration, AI Strategy Roadmap, Agile Development, Scalable ML Systems, Customer-Centric AI Solutions

CERTIFICATIONS

- Google UX Design Certificate (Coursera)
- LangChain for LLM App Development (deeplearning.ai)
- AWS Fundamentals: Core Services

EDUCATION

Master's in Business Analytics — **Sep 2023 – Dec 2024**
Trine University, Michigan, USA

Bachelor's in Engineering — **Jun 2018 – Jun 2021**
Vardhaman college of Engineering, Shamashabad, India

PROJECTS

Autonomous AI Workflow Orchestration for Operations

- Designed agentic AI workflows using LangGraph and CrewAI, automating business tasks and reducing manual handoffs by 40%.
- Integrated Pinecone vector store and Snowflake Cortex for secure RAG pipelines, enhancing system reliability and reducing LLM hallucination rates in production.

GenAI-Powered Support Automation System

- Led multi-agent LLM support deployment, cutting ticket resolution time by 30% and increasing customer satisfaction by 25% across teams.
- Fine-tuned GPT and LLaMA via Vertex AI with CI/CD on GCP, enabling faster iterations and continuous model enhancement cycles.