VAMSHI KRISHNA YENUGULA

New Jersey, USA · 551-375-9713 · yenugulavamshikrishna16@gmail.com · linkedin.com/in/vamshi-krishna-yenugula

Summary

AI/ML Engineer with expertise in **LLM fine-tuning (Mistral-7B, LoRA/QLoRA, 4-bit)** and **Retrieval-Augmented Generation (RAG)** using **LangChain, FAISS, Pinecone, Hugging Face**. Delivered solutions that **reduced migration disruptions by 35% across 300,000+ endpoints**, **lowered outages 20%**, and deployed chatbots with **85%+ accuracy**. Skilled in **MLOps CI/CD**, cloud (Azure, AWS, GCP), and classical ML (XGBoost, Random Forest, AdaBoost, LSTMs). Committed to responsible AI (fairness, bias mitigation, explainability).

Skills

Generative AI	Hugging Face, Transformers, LoRA/QLoRA, 4-bit Quantization, LangChain, Langgraph (Agentic AI), RAG, FAISS, Pinecone, Prompt Engineering, BLEU/ROUGE/Perplexity, OpenAI API, Azure OpenAI
ML/DL	scikit-learn, XGBoost, Random Forest, AdaBoost, PyTorch, TensorFlow, CNNs, LSTMs, Prophet, Isolation Forest, Autoencoders, GANs, Diffusion Models
Data	SQL, PostgreSQL, BigQuery, Pandas, NumPy, Spark, Dask, Kestra, Airflow (familiar), Kafka (familiar)
MLOps	Azure ML, Azure Data Factory, MLflow, Weights & Biases, Docker, Kubernetes, GitHub Actions (CI/CD), AWS S3, GCP AI Platform
Visualization/NLP/CV	Power BI, Tableau, Streamlit, Matplotlib, Seaborn, spaCy, NLTK, OpenCV

Experience

Graduate Research Assistant — Montclair State University   *Sep 2024 – May 2025*

– Built a university QA chatbot that answered 1,500+ queries at **85%+ accuracy**, reducing manual support workload (Mistral-7B LoRA/QLoRA, RAG with LangChain, FAISS, Pinecone, Hugging Face, OpenAI API).
– Deployed a **scalable RAG pipeline** on Docker/Kubernetes, enabling reliable LLM-based Q&A with monitoring in MLflow and Weights & Biases.
– Improved transport KPI predictions by **20%** using time-series forecasting (ARIMA, Prophet) on 10M+ NYC taxi trips with Tableau dashboards for visualization.

Associate Consultant — Capgemini   *May 2023 – Aug 2023*

– Increased migration reliability by **35%** across **300,000+ devices**, saving 100+ hours/week, by building predictive Ensemble ML models (XGBoost, Random Forest, AdaBoost).
– Reduced outages by **20%** through anomaly detection (Isolation Forest, Autoencoders) and reduction of the ticket triage effort **25%** with NLP models (Azure Cognitive Services).
– Developed automation tools (software catalog, BIOS updater, monitoring engine) adopted by **300,000+ employees**.

Senior Analyst — Capgemini   *Oct 2022 – Apr 2023*

– Improved patch compliance by **25%** by developing predictive models using **LSTM networks** on device update logs; enabled infra teams to anticipate failures and remediate issues before impacting **1,000+ endpoints**.
– Cut data prep time by **50%** by building automated **Azure Data Factory + Python pipelines** for 1,000+ endpoints, paired with Power BI dashboards.
– Standardized telemetry schemas, improving consistency of ML feature extraction across systems.

Analyst — Capgemini   *Mar 2021 – Sep 2022*

– Reduced recurring incidents by **20%** by analyzing 20,000+ IT tickets/logs to identify automation opportunities.
– Saved **10+ hours/week** by automating IT tasks (inventory checks, driver validation) with Python; created datasets for downstream ML pipelines.

Education

Montclair State University — M.S. in Data Science   *Expected May 2025*

Relevant coursework: Statistics, Machine Learning, Deep Learning, NLP, Big Data Analytics, Data Mining, Algorithms, Exploratory Data Analysis, Database Systems

Projects

– **Agentic AI with LangGraph:** Built multi-agent workflows to manage reasoning and tool-calling pipelines; improved query handling robustness and reduced latency in RAG-style tasks.
– **Stable Diffusion Image Generation:** Trained Hugging Face Diffusers on a custom dataset; improved FID compared to baseline models.
– **Pneumonia Detection (ResNet-50):** Achieved **93% accuracy** on chest X-rays with explainability via Grad-CAM.
– **Walmart Forecasting:** Improved weekly sales prediction accuracy by **20%** using ensemble methods and Prophet for demand forecasting.

Certifications & Awards

Microsoft Certified: **DP-100**, **AZ-900**, **AI-900** | Rising Star Award (Capgemini) | Process Innovation Nomination