Ramakrishna Vennam

+1 (656) 208-4865 | ramakrishnavennam123@gmail.com | LinkedIn

Summary

AI & Machine Learning Engineer with 4+ years of industry experience building production-grade ML and Generative AI systems across Azure and AWS. Specialized in LLM-based RAG pipelines, multi-agent architectures, computer vision, and MLOps automation. Proven ability to own end-to-end AI solutions, translating complex business problems into scalable, reliable systems in regulated enterprise environments.

Education

University of South Florida
Master of Science in Computer Science — *May 2025*

Bharath University
Bachelor of Technology in Computer Science and Engineering — *May 2021*

Experience

Data Scientist — Jan 2025 – Present
Charles Schwab

- Designed and deployed multi-agent LLM systems utilizing Semantic Kernel and AutoGen, achieving a 98% accuracy rate in automating enterprise knowledge queries, improving response times by 60%
- Engineered end-to-end Retrieval Augmented Generation (RAG) pipelines integrating Azure OpenAI and Azure AI Search, extracting key insights from diverse data sources (PDFs, Word documents, databases)
- Productionized Generative AI services using Docker, Kubernetes, and Azure DevOps, shortening release cycles by 50% while ensuring adherence to governance and auditability standards

Data Scientist — May 2020 – Jul 2023
Accenture — *India*

- Developed and deployed machine learning models (Random Forest, SVM, Decision Trees) for customer analytics and marketing optimization, increasing targeting precision by 10% and driving a 5% lift in conversion rates
- Engineered and managed AWS-based ML pipelines utilizing S3, EC2, Lambda, and SageMaker to enable scalable data preprocessing, model training, and batch inference for over 5 million customer records
- Implemented computer vision and OCR systems with YOLO and PaddleOCR, automating document digitization with 95% accuracy and reducing manual processing time by 70%
- Established CI/CD-driven MLOps workflows with Docker and cloud automation tools, reducing model deployment timelines by 60% and increasing release frequency by 4x

Projects

Production RAG Pipeline for Document Intelligence | *AutoGen, RAG, Azure Functions, FastAPI, LLM's* 2022

- Developed an end-to-end Retrieval Augmented Generation (RAG) pipeline, leveraging AutoGen, Azure Functions, and FastAPI, to facilitate intelligent document processing and semantic retrieval for over 100 users.
- Optimized embedding and retrieval strategies within the RAG pipeline, utilizing LLMs and advanced indexing techniques, which improved query success rate by 25% while ensuring stable and accurate performance across diverse query patterns.
- Implemented a RAG pipeline using AutoGen, Azure Functions, and FastAPI to enable intelligent document processing and semantic retrieval, reducing information access latency by 40% for 100+ users.

Technical Skills

Skills: JavaScript, TypeScript, React, Node.js, Python, Docker, Git, Agile/Scrum, AutoGen, FastAPI, Flask, MLOps, Azure DevOps, GitHub Actions, Azure Web Apps, Azure Functions, ACR, SQL, Generative AI, Machine learning, Deep learning, NLP, Data Visualization, Azure Cloud, Agents, Semantic Kernel, Azure AI Search, RAG, LLMs, Azure VMs

Certifications

AWS Certified Machine Learning – Specialty, Microsoft Azure AI Engineer Associate (AI102), Google Professional Machine Learning Engineer