Sowmiya Velusamy Santhana Krishnan (She/Her)

AI Engineer | Data Scientist

Piscataway, NJ | sowmiya.sign@gmail.com | +1 (469) 417-9470 | https://www.linkedin.com/in/sowmiya-vs/ | GitHub Link

Professional Summary

AI/ML Engineer with 9+ years of IT Industry experience building production-grade generative AI solutions, agentic systems, and scalable ML pipelines. Expertise in developing RAG pipelines, LLM-based applications, and autonomous AI agents using LangChain, CrewAI, and LangGraph, with proven success in deploying models to cloud infrastructure using Google ADK. Effective communicator with a proven ability to collaborate across cross-functional AI and engineering teams to deliver impactful business solutions.

Education

Master of Business and Science: **AI/ML and Data Analytics** | Current GPA: **4/4** Expected graduation **December 2025**
Rutgers University-New Brunswick Campus, NJ, USA
Coursework: AI & Deep Learning, Machine Learning Fundamentals, Business Analytics Programming (Advanced Python), Data Warehousing & Front-End Programming ***(Teaching Assistant)***, Cloud Computing, Data Visualization (Tableau)

Work Experience

AI Engineering Intern | MARV Capital, NY, USA **June 2025 - August 2025**

- Designed and deployed a production-grade AI content moderation system using MCP protocols, achieving 95% accuracy and reducing LLM inference costs by 65% through intelligent escalation workflows
- Employed FAST API endpoints for ML model integration, reducing response latency by 40%
- Optimized prompt engineering strategies for cost efficiency, decreasing cloud spending by $12K annually

Data Scientist Extern (Team Lead) | The Avery Group LLC, NJ, USA **January 2025 - May 2025**

- Led development of AI-powered vendor matchmaking platform using React and MongoDB, improving onboarding efficiency by 60% and serving 500+ vendors
- Built an NLP-based recommendation engine using NAICS code matching, reducing manual review time by 50% and processing 1,000+ vendor profiles
- Delivered data visualization dashboards using Tableau, improving data-driven decision accuracy by 35%
- Coordinated agile sprints and code reviews for a 5-person team, achieving 100% on-time delivery

Data Scientist Extern | Rutgers MBS Food Science Concentration **May 2024 - August 2024**

- Streamlined supervised and unsupervised ML models for toxic contaminant detection, achieving 91% classification accuracy
- Performed statistical modeling and experimental design for food safety analysis, reducing false positives by 25%

Senior Consultant | CAPGEMINI INDIA, Client: NBCU Media **July 2019 - April 2023** ***(4 yrs)***

- Systematized test scripts using Selenium with Java/Python and the TestNG framework, reducing manual testing efforts by 80% and cutting testing cycle time from 45 to 7 days.
- Led multiple product planning, estimations, deliveries, and completed projects 20% faster than planned with zero UAT defects.

Automation & Linux Administrator | IBM INDIA, Client: AT&T Telecom **May 2013 - June 2018** ***(5 yrs)***

- Deployed WAR files and maintained system properties across using shell scripts, and improved deployment efficiency and reduced configuration errors by 50%.
- Implemented Jenkins-based CI/CD pipelines for automated builds and deployments, reducing deployment time by 70%.
- Developed RESTful APIs for a sales application using Java, implementing Maps, Arrays, and core OOP concepts to ensure seamless integration with multiple external systems, improving cross-application communication reliability by 40%.

Skills

Generative AI: Large Language Models (LLMs), GPT, BERT, Transformer Models, Generative AI, RAG Pipelines, Prompt & Context Engineering, LLM Fine-Tuning, Few-Shot/Zero-Shot Prompting, LLM-as-a-Judge, Hallucination Mitigation
Agentic AI Frameworks: LangChain, LangGraph, CrewAI, LlamaIndex, MCP (Model Context Protocol), A2A (Agent-to-Agent), Multi-Agent Systems, Agent Orchestration
Programming Languages: Python (Advanced), SQL, Java, R, Vibe coding (Cursor, Claude Code)

ML/DL Frameworks: PyTorch, Scikit-learn, Hugging Face, NumPy, Pandas, Sentence-Transformers, OpenCV
Cloud Platforms: AWS (EC2, S3, SageMaker, Bedrock), Azure OpenAI, GCP, Vertex AI
MLOps: ML flow, Kubeflow, Docker, Kubernetes, Jenkins, CI/CD Pipelines, Model Monitoring, Model Governance, Observability
Data Engineering: ETL/ELT Pipelines, Apache Spark, Kafka, FAISS, Vector Databases, MongoDB, Data Warehousing
Evaluation & Analytics: Ground Truth Setup, Confusion Metrics, Cost & Latency Monitoring, A/B Testing, Hypothesis Testing, Regression, Forecasting, KPI Tracking, Predictive Modeling, Experimental Design
Soft Skills: Business Acumen, Cross-Functional Collaboration, Strategic Thinking, Leadership

Academic Projects

Make.com Daily Content Automation System

- Developed end-to-end GenAI workflow using OpenAI Assistant API for LinkedIn and Instagram content generation to reduce manual effort by 95%.
- Streamlined a multi-step pipeline with webhook triggers and JSON parsing, processing 30+ content pieces weekly
- Utilized prompt engineering best practices (few-shot, persona-based) to optimize output quality by 50%

Multi-Agent AI Trading Intelligence Platform

- Built a CrewAI-based multi-agent system for financial forecasting using GPT-3.5-turbo and SERP API, improving accuracy by 15%.
- Executed LangGraph and LangChain MCP orchestration for equity screening, market data aggregation, and trade recommendations in a supervised pipeline by improving process efficiency and automation by 90%.

RAG-Based Quantitative Analysis System

- Created an LLM-powered chatbot using OpenAI API and LangChain with 95% precision on context-based queries.
- Achieved 2.5-second average response time using optimized pipeline caching and query chunking.
- Applied hallucination mitigation techniques and observability monitoring for reliability tracking

Computer Vision & Deep Learning Models

- Designed a vehicle detection system using YOLOv8 and YOLOv11 architectures with advanced deep learning frameworks using PyTorch and OpenCV library.
- Applied model fine-tuning with data augmentation techniques, boosting accuracy from 86% to 91%
- Enhanced inference performance for edge deployment, processing 30 FPS with <50ms latency

Achievements

The Avery Group LLC via The MBS Rutgers Exchange Program

- Team Lead Achievement Award
- Best Lightning Talk Presentation Award

Capgemini

- Delivery Excellence Award
- Agile Expertise Award
- Certification for Innovation Idea Implementation Award
- Token of Appreciation – i-SPRINT 2021 – Voice of Change Makers Award

IBM

- Best of IBM Award
- Manager's Choice Award (2014 to 2018)