SASI KUMAR P
AI/ML Engineer | Data Scientist | Generative AI Engineer
Email: sasikumar51825@gmail.com | Mobile: 614-505-5200
LinkedIn: linkedin.com/in/sasikumar99/ | GitHub: github.com/sasikumar51825

PROFESSIONAL SUMMARY

AI/ML Engineer and Generative AI Engineer with over 4 years of experience designing, developing, and deploying end-to-end AI systems with a strong emphasis on Retrieval-Augmented Generation, AI search agents, and production-scale machine learning pipelines. Solid foundation in Computer Science principles with hands-on expertise in Python-based development, machine learning, deep learning, and large language models. Proven ability to build, monitor, and optimize AI pipelines with a focus on accuracy, latency, and reliability in real-world environments.

TECHNICAL SKILLS

Programming Languages: Python, R, SQL
Machine Learning & Deep Learning: Machine Learning, Supervised Learning, Unsupervised Learning, NLP, CNN, RNN, PyTorch, TensorFlow
Generative AI & LLMs: LLMs, GPT, Transformers, MoE, Prompt Engineering, Fine Tuning, RLHF, Ollama
RAG & AI Search: RAG, Data Ingestion, Chunking Strategies, Embeddings, Vector Search, Hybrid Search, LangChain, LangGraph, ReAct, CrewAI, AI Search Agent Architecture
Model Evaluation & Monitoring: Synthetic QA Generation, QA Metrics, Retrieval Hit Rate, Accuracy Metrics, Latency Monitoring, Drift Detection
Data & Analytics: Pandas, NumPy, Tableau
Development & Platforms: GitHub, Kaggle

PROFESSIONAL EXPERIENCE

Associate Data Scientist
Client: BNY Mellon, PA
Mar 2024 – Present
Responsibilities:
• Built end-to-end RAG pipelines covering data ingestion, preprocessing, chunking, embeddings generation, vector and hybrid search
• Designed AI search agents using ReAct reasoning, planning workflows, and multi-agent architectures for complex query resolution
• Implemented grounding and citation mechanisms to trace generated responses back to exact source passages
• Developed automated QA evaluation pipelines using synthetic datasets, retrieval-hit-rate metrics, and continuous feedback loops
• Monitored AI systems in production by tracking latency, accuracy, and performance metrics to ensure reliability at scale
• Orchestrated external tools and enterprise knowledge bases while optimizing inference pipelines for cost and response time
• Collaborated with engineering teams to deploy AI pipelines following best practices in software engineering and system architecture

---

Project Engineer (AI/ML)
Client: Wipro, HYD
Feb 2021 – Jul 2023
Responsibilities:
• Developed machine learning models using Python, PyTorch, and TensorFlow for structured and unstructured datasets
• Built NLP pipelines for text classification, feature extraction, and semantic analysis
• Designed and trained CNN and RNN architectures for predictive modeling and sequence learning tasks
• Applied supervised learning and unsupervised learning techniques to solve data-driven business problems
• Performed data ingestion, cleaning, transformation, and feature engineering using Pandas and NumPy
• Created analytical dashboards and reports using Tableau to support stakeholder decision-making
• Maintained source control and collaborative development workflows using GitHub

---

EDUCATION

Master of Science in Information Technology
Franklin University
Bachelor of Technology in Information Technology
Anil Neerukonda Institute of Technology & Sciences

CERTIFICATIONS

Data Mining with R
Data Visualization with Tableau

PROJECTS

SecureAIChatbot – RAG-Based Secure Generative AI System

Designed and implemented a secure RAG chatbot with vector embeddings, hybrid search, and semantic retrieval. Integrated AI grounding and citation to reduce hallucinations and improve trust. Applied Red Teaming techniques and evaluation using synthetic QA datasets and accuracy metrics. Deployed LLM integration using Ollama with optimized prompt engineering and end-to-end pipeline monitoring.

VoiceBot – AI-Powered Interview Assistant

Built a voice-based interview assistant using Whisper for speech-to-text and GPT-based models for intelligent feedback. Designed end-to-end AI pipelines with offline inference, latency optimization, and customizable question flows while ensuring privacy and model reliability.