SAI MANIKANTA KASIREDDY

Machine Learning Engineer

(940) 465-1437 | ksaimanikanta4@gmail.com

United States| LinkedIn | GitHub | LeetCode | HackerRank | Medium

SUMMARY

AI Engineer with 4+ years designing and deploying enterprise-grade GenAI systems, multi-agent workflows, and RAG-driven architectures in cloud-native environments. Deployed HIPAA-compliant healthcare RAG platforms on AWS Bedrock (92% recall) and built agentic copilots using Claude, LangGraph, and Model Context Protocol (MCP) integrated with Jira and GitLab. Experience spans Azure (AKS, Databricks, Data Factory, DevOps) and AWS (Bedrock, SageMaker, Lambda) with strong Python, C#, and JavaScript skills. Co-author ACL 2024 in healthcare AI, LeetCode 500+, HackerRank 5-star Python.

TECHNICAL SKILLS

Domain Expertise: Multi-Agent System Design, Microsoft Azure AI Foundry (patterns) ,Copilot / Agent-Based Systems · Semantic Kernel (concepts), RAG Pipelines & RAGAS Evaluation, HIPAA-Compliant Data Architecture, Healthcare AI & Medical Chatbots, AWS Bedrock Agents, MCP / A2A, Enterprise Integrations (Salesforce, Jira, GitLab, MongoDB, Stripe)

Languages: Python, R, C++, SQL, Java, C#, JavaScript

ML/DL: PyTorch, TensorFlow, Keras, Scikit-Learn, XGBoost, Transformers (GPT, LLaMA, Mistral), Fine-Tuning, PEFT, LoRA, QLoRA

GenAI & LLMs: LangChain, LangGraph, LlamaIndex, RAG, Agentic AI, Prompt Engineering, Hugging Face, Ollama, OpenAI, Anthropic Claude

Cloud & MLOps: Azure(AKS, Databricks, Data Factory, DevOps), Azure AI Foundry(patterns),AWS (SageMaker, Bedrock, Lambda, S3, Glue, Athena, EMR), GCP, MLflow, Docker, Kubernetes, CI/CD

Data & Tools: Spark, Airflow, FAISS, Pinecone, PostgreSQL, MongoDB, DynamoDB, FastAPI, Flask, Tableau, Power BI, Git, N8N,

EXPERIENCE

ML Research Scientist — Multi-Agent VLMs & Foundation Models | Auburn University, Prof. Pan He (R&D) + Harbert College of Business *Jan 2026-Present*

- Investigating multi-agent coordination strategies for Vision-Language Models (VLMs) applied to industrial anomaly detection, prototyping orchestration patterns (AutoGen, chain-of-thought, tool calling) analogous to Azure AI Foundry and Copilot-style multi-agent workflows.
- Built production ML data pipelines, personalization classification algorithms, and automated evaluation frameworks on behavioral telemetry for 200+ portfolio companies owning full ML lifecycle from experimentation through deployment and monitoring
- Deployed privacy-preserving AI analytics using Anthropic Claude and LangGraph with on-device processing capabilities; conducted code reviews enforcing best practices for maintainable, extensible ML code partnering cross-functionally with faculty, founders, and staff as a trusted ML science partner.
- Designed and maintained scalable data infrastructure with governance standards, quality checks, and Salesforce CRM pipelines; integrated telemetry and instrumentation tracking system performance, user behavior, and operational metrics

Hantec Global Inc. | AI/ML Engineer(Data Scientist) *Jan 2025 – Dec 2025*

- Built and launched two production GenAI copilot-style solutions for major electric delivery client: automated Jira workflow system using Claude with Model Context Protocol, and intelligent PR summary generator for GitLab, both fully productionized on AWS (Bedrock, SageMaker, Lambda, S3, DynamoDB, QuickSight) streamlining operations and improving efficiency by 40%.
- Experimented with Infrastructure-as-Code tools including AWS CloudFormation and Terraform for provisioning, Kubernetes for container orchestration, and AWS CodePipeline for automating model deployment workflows reducing deployment time by 45%
- Designed and delivered enterprise-grade LLM-Ops framework on AWS building GenAI reference architecture with S3, Lambda, Bedrock Knowledge Bases, Amazon Titan embeddings, and OpenSearch powering RAG pipelines with built-in guardrails, versioning, and monitoring achieving 92% recall rate on domain-specific queries, using patterns directly portable to Azure AI Foundry RAG solutions.
- Implemented HIPAA-compliant data lake with AWS Glue triggers, schedulers, Lambda functions, Athena dashboards, DynamoDB, and S3 raw/processed buckets, standardizing ingestion, schema evolution, and anomaly detection for healthcare data processing
- Automated reporting and analytics using Athena + QuickSight dashboards cutting manual reporting time by 50%, while improving pipeline reliability through CI/CD-enabled ETL orchestration for repeatable, client-ready deployments
- Implemented Amazon Bedrock services and Bedrock Agent to integrate foundation models (Claude, Titan) into client workflows, enabling secure, scalable, and agentic copilot-like applications with multi-step reasoning and tool use capabilities.

University of North Texas | ML Research Engineer(Research Associate). *Jan 2023 - Dec 2024*

- Led cross-functional COVID-19 research study building data annotation workflows and ingestion pipelines, uncovering data quality issues through anomaly detection, resolving schema mismatches, enhancing insight accuracy, ensuring compliance, and accelerating research timelines by 20%
- Spearheaded development of AI-driven predictive models using LangChain, LangGraph, and LlamaIndex, conducting comprehensive data mapping, modeling, and documentation to enhance traceability, schema validation, and alignment with evolving healthcare business needs
- Engineered AI-powered medical assistant chatbot leveraging Retrieval-Augmented Generation (RAG) approach with Ollama, LangChain, FAISS, and agentic AI workflows via LangGraph, functioning as a healthcare copilot and achieving 25% increase in user engagement and 92% recall rate of essential medical information
- Established structured evaluation framework incorporating RAGAS metrics (faithfulness, context relevance, answer relevance) to measure chatbot performance, reduce fallback rates by 15%, and guide iterative improvements enhancing user satisfaction and trust in healthcare guidance
- Applied advanced data mining and statistical modeling techniques to identify healthcare risk factors and patient outcomes, collaborating on GenAI and knowledge graph implementations for regulatory compliance, risk assessment, and ethical AI integration

- Co-authored ACL 2024 paper "Altriva - AI-Powered Chatbot for Personalized Alternative Medicine through GenAI" top-tier NLP/AI publication

Tata Consultancy Services | Machine Learning Engineer *Jan 2021 - Dec 2022*

- Spearheaded design, development, and deployment of machine learning models for Vodafone UK (10M+ users) leveraging Python and industry-standard ML frameworks (Scikit-Learn, XGBoost, TensorFlow) boosting churn prediction accuracy by 20% and reducing customer unsubscribes by 15%
- Led comprehensive business analysis and ML modeling using logistic regression, decision trees, random forests, and ensemble methods to identify key churn drivers, improve customer segmentation, and optimize resource allocation boosting retention and operational efficiency by 30%
- Streamlined end-to-end telecom data processing by implementing CI/CD pipelines using Azure DevOps and optimizing ETL workflows with Azure Data Factory, Databricks, and Spark (Scala), reducing production load times by 40% and batch processing times by 30% using Azure-native pipelines that mirror enterprise healthcare data-processing stacks.

FEATURED PROJECT

Project Silo - Private On-Device AI on Apple Silicon

- Built fully offline, privacy-preserving AI agent on Apple Silicon (ANE) using MLX-quantized Llama-3.2-1B achieving 125 tokens/sec with less than 1.1GB peak RAM featuring on-device RAG (NaturalLanguage embeddings), thermal-aware inference control, and lightweight SwiftUI macOS Menu Bar interface
- Developed native macOS application using Swift/SwiftUI showcasing deep expertise in Apple ecosystem development and Apple Human Interface Guidelines
- Implemented on-device RAG using Apple's NaturalLanguage framework eliminating cloud dependencies and ensuring complete user privacy with zero data transmission
- Optimized for Apple Neural Engine (ANE) achieving real-time performance with minimal battery impact through thermal-aware inference control and Metal Performance Shaders

MovieLLM - Fine-Tuned LLM for Recommendations | github.com/saikasireddy/moviellm

- Fine-tuned LLaMA 2 (7B) using LoRA/QLoRA achieving 87% accuracy with 4-bit quantization, optimized data structures and algorithms for efficient inference
- Built production system using C++/Python for low-latency serving, deployed FastAPI REST API serving 1000+ req/min with 200ms latency, demonstrated 35% improvement through A/B testing
- Implemented comprehensive automated testing framework with unit tests and regression tests achieving 95% code coverage
- Designed with privacy-conscious architecture allowing optional on-device deployment for sensitive use cases

FetchMe — AI-Powered Restaurant Owner Dashboard (Agentic Prototype) In Development – Mar 2026

- Architecting and building an AI-powered owner dashboard prototype for FetchMe, a live restaurant marketplace platform processing $3M+ annual revenue across hundreds of restaurant clients — integrating Claude (Anthropic Bedrock) and agentic workflows via LangGraph/N8N with FetchMe's production NextJS + MongoDB backend
- Designing multi-path RAG architecture across three MongoDB deployment configurations (Atlas M10+ with native Bedrock Vector Search connector, Atlas free-tier with pymongo naive retrieval, and self-hosted bridge via S3-sync Lambda pipeline) enabling restaurant owners to query sales data, identify top customers, update menus, and generate marketing insights through natural language chat
- Prototype scope: conversational dashboard interface covering 4 core owner workflows menu/price updates, most valuable customer identification, revenue reconciliation against Stripe payments, and best-selling item analysis for marketing campaigns; voice ordering integration planned as Phase 2

EDUCATION

Master of Science in Data Science | University of North Texas *Jan 2023 - Dec 2024*

Bachelor of Technology in Electronics & Communication | Pragati Engineering College, India *Jun 2017 - Jun 2021*

PUBLICATIONS & ACHIEVEMENTS

- **Co-author, ACL 2024: "Altriva - AI-Powered Chatbot for Personalized Alternative Medicine through GenAI"**
- **Co-author:** "An Approach for Ensuring Privacy in Smart Contracts through GenAI"
- **LeetCode:** 500+ problems solved across Data Structures, Algorithms, Dynamic Programming, and System Design
- **HackerRank:** 5-star Python certification, 4-star Problem Solving | Top 5% in algorithms competitions
- **GitHub:** 20+ open-source ML projects with 500+ stars | Active contributor to LangChain, Hugging Face, and PyTorch communities
- **Medium:** Published technical articles on LLM fine-tuning, RAG architectures, and production ML systems