Aasritha Sakhamuri

AI/ML Engineer | Generative AI | NLP | Deep Learning | LLM | MLOps

Overland Park, Kansas | (913) 334-7545 | aasrithas83@gmail.com | LinkedIn | GitHub

SUMMARY

AI/ML Engineer with 3+ years of experience designing and deploying advanced machine learning and generative AI systems across healthcare and enterprise environments. Specialized in building LLM-powered diagnostic tools, reinforcement learning agents, and multimodal sentiment analysis pipelines using GPT-4, LangChain, and HMMs. Proven success in developing scalable MLOps platforms, integrating vector databases (FAISS, Pinecone), and optimizing model performance in production. Adept at translating complex AI techniques into real-world impact—boosting clinical accuracy, automating feedback loops, and enhancing patient engagement. Cloud-native practitioner with hands-on expertise in Azure, Docker, Kubernetes, and end-to-end model lifecycle management.

SKILLS

Languages: Python, SQL, Bash, Java (8–21), C++
Machine Learning & Deep Learning: XGBoost, Random Forest, CNN, LSTM, RNN, Autoencoders, Grad-CAM, TensorFlow, PyTorch, Scikit-learn, NumPy, Pandas, Keras, FastAI, Spark
NLP Tools: spaCy, Transformers, BERT, Regex, Entity Recognition, BERTScore, ROUGE-L
Generative AI & LLMs: GPT-4, Gemini, Claude, Llama, BERT, T5, Hugging Face Transformers, LangChain, RAG
Fine-Tuning & Optimization: LoRA, QLoRA, PEFT, Prompt Tuning, Few-Shot Prompting, System Prompts, Quantization
Reinforcement Learning: PPO, DDPG, SAC, Bandit Algorithms (UCB, Thompson Sampling), RLHF, DPO
Vector Search & RAG Systems: FAISS, Pinecone, ChromaDB, LlamaIndex, Chunking Strategies, Vector Indexing
MLOps & Model Development: MLflow, Apache Airflow, Docker, Kubernetes, GitHub Actions, REST APIs, FastAPI, Gradio
Monitoring & Observability: Prometheus, Grafana, Weights & Biases
Big Data & Distributed Computing: Spark, PySpark
Cloud Platforms: AWS (SageMaker, Lambda, S3), Azure (AKS, Cognitive Search), GCP (Vertex AI, Storage)
Databases: PostgreSQL, MySQL, MongoDB
Testing & CI/CD: Pytest, Model Validation Pipelines, CI Test Coverage
Data Visualization & UI Development: Matplotlib, Power BI, Streamlit
Version Controlling: Git, GitHub

EXPERIENCE

AI Engineer | CitiusTech, USA **Jan 2025 – Present**

- Led the development of an LLM-based clinical reasoning assistant using **GPT-4**, **Gemini**, **LangChain**, **LlamaIndex**, **FAISS**, and **Azure Cognitive Search**, cutting diagnostic turnaround time by 40% and improving care team coordination.
- Designed and deployed patient risk prediction models using **LSTM reinforcement learning agents** and **Hidden Markov Models**, resulting in a 25% increase in clinical outcome accuracy and treatment optimization.
- Built multilingual **conversational AI** interfaces by integrating **OpenAI embeddings**, **Pinecone vector search**, and custom **intent detection models**, boosting patient portal engagement by 3x.
- Automated patient feedback workflows using **LSTM Autoencoders** and **RLHF-style fine-tuning**, achieving a 37% improvement in review clustering and triage response time.
- Developed a **real-time anomaly detection engine** for healthcare transactions using **semantic graph embeddings** and classification layers, reaching 92% precision and reducing processing latency by half.
- Created a **multimodal sentiment analysis** system leveraging **FastText**, **spaCy**, **LSTM**, and **HMMs** to analyze patient voice, video, and chat data with over 90% F1-score for intent recognition.
- Implemented robust **MLOps pipeline** using **FastAPI**, **Docker**, **MLflow**, **Azure Kubernetes Service**, **Airflow**, **Prometheus**, and **GitHub Actions** for continuous model training, vector index updates, and system monitoring.

Machine Learning Engineer | Streebo Inc **Dec 2020 – July 2023**

- Developed a **transformer-based clinical forecasting system** by integrating **SQL-based ETL pipelines** with geospatial **data fusion**, enabling accurate predication of patient volumes and resource needs – reducing planning errors by 12% across regional hospitals.
- Engineered a **spatiotemporal inventory optimization** model using **Graph Attention Networks**, **TensorFlow**, **MongoDB**, and **Apache Airflow**, reducing stockouts in critical supply categories by 19%.
- Designed an **intelligent patient routing engine** for mobile health fleets using **reinforcement learning algorithms** (PPO, DDPG) and **graph search**, minimizing transportation costs by 28% while improving access in underserved regions.
- Implemented a real-time **imaging AI pipeline** using **FastAPI**, **AKS**, and **asynchronous data streaming** to process satellite and drone imagery for epidemiological surveillance in under 5 seconds.
- **Applied Bayesian uplift modeling** and counterfactual **tree-based classifiers** to personalize preventive outreach strategies, resulting in a 21% uplift in public health campaign engagement.

- Built a **GenAI powered** virtual assistant using **GPT-4**, **LangChain**, **ChromaDB**, and **vector retrieval** to deliver energy-efficiency recommendations to hospitals – boosting sustainability program participation by 34%.
- Established an end-to-end **MLOps pipeline** for **GenAI** workloads using **MLflow**, **GitHub Actions**, **Docker**, and **Kubernetes**, supporting secure CI/CD workflows, model evaluation, and scalable deployment of healthcare AI models.

EDUCATION

Master of Science in Computer Science | University of Central Missouri | Lee's Summit, Mo **May 2025**
Bachelor of Technology in Information Technology | VVIT | Guntur, India **April 2023**

PROJECTS

AI-Powered Virtual Assistant | Link

An intelligent voice-enabled assistant built using **Rasa**, **Google Speech-to-Text**, **BERT**, and **FastAPI** for:

- Context-aware conversation memory via Rasa Tracker; voice-based task scheduling using Google Calendar API
- Real-time information retrieval from Wikipedia and Weather APIs; voice I/O through **PyAudio** and **pyttsx3**
- Modular backend architecture with **Flask**; deployment on Google Cloud (planned)

Multi-Modal RAG System for Voice, Image, and Document Q&A | Link

A full-stack retrieval-augmented generation system built with **GPT-4o**, **FAISS**, and **Streamlit** to answer questions from PDFs, images, and audio files.

- Extracts content using **PyMuPDF** (PDF), **Tesseract OCR** (images), and **Whisper** (audio); embeds with **OpenAI ADA-002**
- Performs semantic search with **FAISS** and generates grounded answers using **GPT-4o** with inline citations
- Features a modular Python backend and ocean-themed **Streamlit UI**; organized, Git-tracked, and packaged for deployment

LLM-Powered RAG System for Internal Knowledge Search | Link

A full-stack GenAI application built with **GPT-4o**, **FAISS**, and **FastAPI + Streamlit** to enable natural language Q&A over internal documents.

- Implemented document parsing with **PyMuPDF** and chunking via **LangChain splitters**; embedded text using **OpenAI ADA-002**
- Performed semantic search using **FAISS**; generated grounded, cited answers with **GPT-4o** using role-based prompts
- Built a modular backend (FastAPI) and chat-style frontend (Streamlit) with inline citations, source chunk expansion, and user feedback collection
- Packaged with **venv**, tested with **unit scripts**, and deployed on **GitHub** for demo and reuse

PUBLICATION

Live Capturing Based Image Segmentation Using Mask R-CNN | Link

International Journal (Volume 12, Issue 4, April 2023)

Built a real-time instance segmentation model using **Mask R-CNN** for live video-based multi-object detection and pixel-level classification.

- Enabled accurate segmentation from streaming input using **OpenCV**, CUDA-accelerated inference, and pre-trained COCO weights
- Contributed to model architecture tuning, custom dataset annotation, and performance evaluation using IoU and mAP metrics
- Co-authored publication; led experimentation pipeline, hyperparameter tuning, and result visualization

CERTIFICATIONS

- Microsoft Certified – Azure AI Engineer Associate (Microsoft)
- AWS Certified Developer – Associate (AWS)
- Microsoft Certified - Azure Fundamentals (Microsoft)
- Prompt Engineering for Developers (DeepLearning.AI)
- LangChain for LLM Application Development (DeepLearning.AI)
- Advanced Generative AI for Developers (Google Cloud)