SATHWIK KUCHANA

+1 (201) 705-9409 | kssathwik@outlook.com | Jersey City, NJ, USA | linkedin.com/in/sathwik-kuchana/

PROFESSIONAL SUMMARY

AI Engineer with 2+ years of experience designing and deploying production-grade AI systems, specializing in LLMs, multi-agent architectures, and RAG pipelines. Expert in fine-tuning transformer models, building scalable AI-driven platforms on AWS, and leading cross functional teams to deliver enterprise AI solutions in healthcare and fintech domains

PROFESSIONAL EXPERIENCE

Valuai.io **San Diego, CA, USA**
AI Engineer *April 2025 - Present*

- Architected and deployed an AI-driven clinician–patient engagement platform using AWS Bedrock, LangChain, FastAPI, Redis, and MongoDB, ensuring scalable, and secure healthcare interactions
- Designed multi-agent orchestration workflows with retrieval-augmented generation (RAG) powered by evidence-based medical literature and clinical guidelines, delivering personalized care plans and reducing clinician response time
- Fine-tuned large language models (LLaMA 3.1 8B) using synthetic datasets of patient–clinician conversations; deployed fine-tuned models on Amazon SageMaker JumpStart and Fal for production inference
- Built multi-persona conversational LLM agents with dynamic routing based on user inputs, optimizing latency, cost, and response quality across diverse patient demographics and clinical intents
- Developed a knowledge base by parsing and structuring high-quality PubMed papers and clinical guidelines into JSON, integrating them into a custom RAG pipeline to ensure output accuracy and explainability
- Led a team of AI interns working on voice model integration (TTS/STT), agent tooling, fine-tuning pipelines, and LangChain-based orchestration; conducted regular code reviews, roadmap planning, and performance mentoring
- Automated generation of 20,000+ synthetic patient profiles using schema-guided combinations of substance use, psychiatric comorbidities, pain subtypes, and demographics to support large-scale training and evaluation
- Built real-time monitoring and alerting pipelines to detect and notify clinicians of critical patient signals, enabling proactive, data-driven interventions and reducing response latency
- Developed and maintained a containerized FastAPI backend, deployed on AWS EC2 with Docker, supporting real-time text and voice interactions, authentication via AWS Cognito, and Redis-powered session memory
- Launched a SwiftUI-based iOS application enabling secure, voice-based patient interaction, tightly integrated with backend AI agents and real-time LLM inference services.Engineered advanced prompt strategies (few-shot, zero-shot, chain-of-thought, persona-driven) to improve LLM accuracy, reasoning quality, and clinical safety.Implemented role-guided prompt routing and safety guardrails to minimize hallucinations and maintain standardized decision support in healthcare workflows.

Nixacom **North Carolina, USA**
Artificial Intelligence Development Intern *March 2025 - May 2025*

- Leverage the Crew AI Framework to automate finance application processing by integrating advanced vision LLM tools for extracting critical data from customer applications, IDs, paystubs, selfies, and 30-second videos
- Engineer robust data extraction pipelines using LLMs to accurately parse and verify information from diverse document types, enhancing data integrity and processing efficiency
- Implement real-time liveness detection using OpenCV, MediaPipe, and DeepFace to strengthen security and prevent identity fraud
- Build and optimize LLM-based agents for automated decision-making, cross-referencing data with the employee database to determine precise loan eligibility.

Jocata Financial Advisory & Technology **Hyderabad, Telangana, India**
Software Engineer *September 2021 - August 2023*

- Developed and fine-tuned LLMs using advanced models like GPT, BERT, and Hugging Face, enhancing conversational AI capabilities and optimizing performance for real-time, dynamic applications
- Integrated RAG workflows leveraging Pinecone and other vector databases to enable scalable, domain-specific knowledge retrieval, enhancing decision-making for enterprise systems
- Built LangChain-based applications, utilizing OpenAI and Hugging Face APIs to streamline document processing and automate natural language workflows, boosting operational efficiency
- Designed and maintained CI/CD pipelines using Jenkins and GitLab, improving deployment efficiency, reducing release cycles by 30%, and ensuring automated testing and integration for machine learning models
- Collaborated with cross-functional teams and AWS services like Lambda, S3, and EC2 to implement AI-driven solutions, ensuring seamless deployment and integration of scalable, cloud-based ecosystems.

Vectra Automation, Inc **Chennai, Tamil Nadu, India**
Software Engineer Intern *April 2021 - June 2021*

EDUCATION

Yeshiva University **September 2023 - May 2025**
Master's, Artificial Intelligence *GPA: 3.8*

PROJECTS & OUTSIDE EXPERIENCE

Climate Change Evidence Synthesis Using LLMs

- Leverage open source LLM models (e.g., Llama, Falcon, Mistral) for synthesizing climate change evidence from scientific literature.
- Apply advanced fine-tuning techniques, including LoRA, QLoRA, and model quantization, to boost model performance.
- Prepare custom, model-specific datasets and design tailored prompt strategies for optimized information extraction.
- Continuously refine accuracy for three key tasks: geographic location extraction, stakeholder identification, and adaptation depth classification to drive systematic, scalable climate change assessments.

Automation of Vertebral Heart Size Detection in Canine Radiographs Using Deep Learning Frameworks

- Designed a hybrid deep learning model integrating Vision Transformers (ViT) and CNNs for automated vertebral heart size detection, achieving a test accuracy of 79%.
- Implemented geometric validation layers to enforce orthogonality, enhancing measurement precision and diagnostic reliability.
- Curated and processed a dataset of 2,000 canine radiographs with advanced augmentation techniques to improve model generalization and accuracy.

Enhanced Image Segmentation Techniques for Bird Sound Spectrogram Analysis

- Developed a custom CNN achieving 64.41% IoU and 78.35% F1 Score in bird sound spectrogram analysis.
- Optimized data preprocessing with advanced techniques for noise reduction in spectrograms.
- Improved bird call identification to support ecological monitoring and conservation efforts.

Fraud Detection in Health Care Charges

- Applied PCA for dimensionality reduction and KNN classification to detect fraudulent healthcare claims, enhancing model accuracy and interpretability.
- Engineered domain-specific features like ratios and logs, improving fraud detection precision through advanced feature engineering.
- Automated data preprocessing, including outlier detection and handling missing values, ensuring robust input for machine learning models.

Object Detection

- Prepared Custom dataset using MNIST Hand written digits dataset by placing each digit image on a black canvas of width 75 x 75 at random locations
- Calculated mean IOU for evaluating the models performance
- Used custom CNN model with regression to perform classification and object localization

SKILLS

Programming Languages: C, C++, Java, JavaScript, Python
Frameworks / Libraries: LangChain, LangGraph, CrewAI, TensorFlow, PyTorch, Keras, FastAPI, NumPy, Pandas, ROS2
Databases: Redis MemoryDB, MongoDB, FAISS, PostgreSQL, MySQL
Cloud / DevOps: AWS Bedrock, Amazon SageMaker, AWS EC2, AWS S3, AWS Lambda, AWS Cognito, AWS CloudWatch, Azure Cloud, Docker, GitHub Actions
Tools / Analytics: OpenAI APIs, Prompt Engineering, Guardrails, JSON Mode, Function Calling, Structured Outputs, n8n Automation, Document Chunking, Document Parsing
Other:Generative AI, AI Agents, Agentic AI, Multi-Agent Systems, RAG Pipelines, Knowledge Bases, Embeddings, Vector Databases, Knowledge Graphs, Fine-Tuning, LoRA, QLoRA, Serverless Architectures, Microservices, Async Workflows, Microsoft Copilot, Azure AI Studio, Microsoft Foundry

CERTIFICATIONS

AWS Academy Graduate - Cloud Architecting
AWS Academy Graduate - Cloud Foundations
AWS Academy Graduate - AWS Academy Data Engineering
Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning