

Kalyanam Venkata Sree Sai

Applied AI Research Engineer



kvsreesai.ai.engineer@gmail.com



2055789980



Birmingham, AL



Summary

Research-driven Applied AI Engineer specializing in **Generative AI, Large Language Models (LLMs), and Multimodal Systems**. Proven track record in bridging the gap between academic research and production-ready AI applications. Experienced in fine-tuning models, designing RAG pipelines, and building scalable AI microservices.



Education

Master of Science in Computer Science, University of Alabama at Birmingham

08/2024 – 05/2026 | Birmingham, AL

Focus: Machine Learning, Deep Learning, NLP, Computer Vision

Bachelor of Technology in Computer Science (AI), Koneru Lakshmaiah Education Foundation

08/2020 – 05/2024 | Guntur, AP

- GPA: 8.6 / 10 | Focus: Artificial Intelligence, Research



Skills

AI/ML

- Py-Torch.
- TensorFlow.
- Scikit-learn.
- Reinforcement Learning.
- Neural Networks

Generative AI

- LLM Fine-Tuning.
- RAG (Retrieval-Augmented Generation)
- Prompt Engineering.
-

Backend & Infrastructure

- FastAP.
- Node.js.
- Docker.
- AWS.
- REST APIs

Languages & Tools

- Python.
- JavaScript.
- React.
- Next.js.
- Git.

Computer Vision

- OpenCV.
- MediaPipe.
- Object Detection.
- Image Segmentation



Certificates

AWS

- Solutions Architect – Associate.
- Developer – Associate.
- Cloud Practitioner.

Specialized

- TensorFlow Developer Certificate.
- Python & CV Bootcamp

Azure

- AI-900.
- AZ-900.
- DP-900.



Publications

Machine Learning and Deep Learning Techniques for Coronary Heart Disease Prediction, Published in IEEE ICCMC 2023 (February 2023)

02/2023

- Conducted comparative analysis of traditional ML models (SVM, Random Forest) vs. Deep Learning (ANN).
- Engineered advanced feature selection techniques, achieving a peak accuracy of **84.44% with ANN**.

- Demonstrated the scalability of AI-driven clinical decision support systems.

Projects

Virtual AI Try-On System, Multimodal AI Perception & Recommendation Engine

04/2026 – 05/2026

- Designed a real-time AI pipeline integrating **Computer Vision** and **Recommendation Engines**.
- Implemented body detection and feature extraction to generate high-fidelity digital avatars.
- Built a custom engine for virtual garment overlay, reducing retail return rates through improved size confidence.

AI Career Copilot, RAG-Based Intelligent Resume & Job Matching Assistant

05/2026 – 05/2026

Architected a scalable **Retrieval-Augmented Generation (RAG)** pipeline to automate resume analysis and career coaching.

- Optimized **Large Language Model (LLM)** responses, resulting in a **35% increase in job-matching accuracy** and a significant reduction in screening time.
- Built a production-ready interface focusing on seamless user interaction and high-performance data retrieval.
- **Tech Stack:** LLM, Next.js, and Vector Databases.

AI Microservices Platform, High-Performance Scalable AI API System

05/2026 – 06/2026

- Developed a containerized microservices architecture using **Docker** and **FastAPI** to handle high-frequency AI inference requests.
- Successfully managed over **10,000+ requests** while reducing overall system latency by **40%**.
- Implemented modular system design principles to ensure horizontal scalability across **AWS** infrastructure.
- **Tech Stack:** Python, FastAPI, Docker, and AWS.

LLM Infrastructure & Adaptive Quantization, High-Throughput Inference Optimization

06/2026 – 08/2026

- Designed a custom inference layer to deploy Large Language Models (LLMs) on resource-constrained environments using **Advanced Quantization** techniques.
- Implemented **KV-Caching** and **Continuous Batching** to improve token throughput by **2.5x** compared to standard baseline deployments.
- Developed an automated evaluation harness to monitor model drift and hallucination rates in real-time production environments.
- **Tech Stack:** PyTorch, vLLM, Docker, and NVIDIA Triton Inference Server.

Interests

Generative AI Research

Exploring the boundaries of Large Language Models (LLMs), specifically in **Retrieval-Augmented Generation (RAG)** and model fine-tuning

Low-Code Innovation

Experimenting with low-code platforms like **Mendix** and **Zoho** to rapidly prototype and deploy AI-driven business solutions

AI Education & Content Creation

Developing a YouTube channel, "**Novarion AI**," dedicated to teaching Artificial Intelligence from zero to expert levels

Technical Research

Investigating the practical application of **Machine Learning** in high-impact fields like healthcare and predictive diagnostics.

Multimodal Systems

Designing intelligent applications that bridge the gap between **Computer Vision** and natural language processing.

Indian Cinema & Cultural Tech

Following the evolution of the **Telugu film industry** and exploring how AI can be applied to media and entertainment.