

ARYAN MAHESHWARI

Los Angeles, California | (213) 272-9648 | aamahesh@usc.edu | [LinkedIn](#) | [Personal Website](#) | [GitHub](#)

EDUCATION

University of Southern California, MS in Applied Data Science | Los Angeles, California

December 2026

K.J Somaiya Institute of Technology, BTech in Artificial Intelligence and Data Science | Mumbai, India

May 2024

TECHNICAL SKILLS

Languages: Python, Java, C++, Railway, ONNX, SQL, linux, pandas, TensorFlow, TensorRT, Next.js, Vercel, pytorch, Scikit-Learn

Machine Learning: Diffusion Models, LLM Fine-Tuning, Deep Learning, Computer vision, RAG, Vector Search, Slack API

ML Systems/Infrastructure: LLM Evaluation, Langchain, Langgraph, LLM Fine-tuning, CUDA, AutoGen, N8n, Zapier, MLflow

Cloud & DevOps: AWS (S3, OpenSearch), GCP (Vertex AI, Dialogflow), Docker, CI/CD, git

PROFESSIONAL EXPERIENCE

USC Marshall School of Business ([GitHub](#))

Los Angeles, California

AI Engineer

December 2025-Present

- Engineered **Large Language Model (LLM)** based ETL pipelines and agentic **Natural Language Processing (NLP)** workflows to **semantically** index **500+** documents across **10+** sources, enabling structured policy data analysis
- Improved structured document **extraction** accuracy by **25%** and reduced document parsing errors by **35%** by developing a **multimodal Vision-Language Model (VLM)** and **OCR pipeline** for PDFs, scans, and tables.

Convexia (YC'25) ([GitHub](#))

San Francisco, California

AI Engineer

March 2025-September 2025

- Built end-to-end **evaluation** pipeline integrating **6** classification **machine learning** models with **MLflow** tracking, **SHAP**-based feature **visualization**, and confidence detection across biomedical and clinical datasets, achieving **100% reproducibility**
- Reduced **CI/CD** runtime by **30%** by **modularizing training and inference** workflows, implementing structured **logging**, and automating environment setup across the **ML infrastructure**

Easley Dunn Productions, Inc

Los Angeles, California

Machine Learning Engineer

June 2025-September 2025

- Improved player-coach match accuracy by **25%** and reduced class imbalance by **40%** by building a **hybrid ML recommendation engine** leveraging **15+** features with rule-based constraints
- Led cross-functional team of **8** engineers to design, build, and deploy an AI-powered sports simulation and **personalization** platform, overseeing **model architecture**, **data pipelines**, and **production deployment**

USC Autodrive Lab

Los Angeles, California

Machine Learning Engineer

June 2025-Present

- Engineered **perception**, **motion prediction**, and **planning** models for **end to end autonomous** driving; deployed **transformer-based generative AI** on **NVIDIA CUDA** clusters, achieving **4x** faster training throughput
- Developed and trained **Reinforcement Learning (RL) algorithms (PPO, SAC)** for autonomous vehicle navigation, achieving **0.4 m** mean positional error across **500+** closed-loop test runs

AGIE AI

Mumbai, India

AI Engineer

January 2024-November 2024

- Reduced chatbot response **latency** by **23%** (**1.3s** to **1.0s**) by integrating **Conversational AI** using **Dialogflow** with **Vertex AI** on **Google Cloud Platform (GCP)**, improving user engagement across **2** pilot client campaigns
- Built an **embedding-based retrieval and ranking pipeline** using **cosine similarity**, achieving **80%+ precision@k** in relevance matching for high-potential startup discovery

Dawn Digitech

Mumbai, India

AI Engineer

January 2023-December 2023

- Designed **Large Language Model (LLM)** based pipelines for AI-driven interview systems, implementing **sentiment analysis** and candidate **evaluation workflows** using **large language models** and **open-source** frameworks
- Improved sentiment **classification accuracy** by **20%** by building a hybrid analysis system combining **BERT**, **VADER**, and **LLM-based prompting**, enabling **context-aware** interpretation across diverse inputs

PROJECTS

EduMate.ai ([GitHub](#))

- Engineered an **agentic AI**-powered educational platform with **RAG-based Q&A**, quiz generation, and **real-time** chat using **Cohere** embeddings, **Pinecone**, **LangChain**, **GPT-4**, and **FastAPI**, supporting **semantic** retrieval across **10,000+** pages of textbook content

Inferno ([GitHub](#))

- Built Inferno, an **LLM inference optimization engine** implementing **INT8 KV cache quantization** (per-tensor and per-channel) and a **continuous batching** scheduler from scratch on **HuggingFace Transformers**; **benchmarked** on **Tesla T4 GPU** achieving **2x** KV cache **compression** with **1.58 PPL** degradation, all **46** tests passing

ACHIEVEMENTS

- Won **1st** place among **100+** teams at MiroxKiro LA Hackathon by architecting **multi-agent AI** system with **LLM** based reasoning, retrieval pipelines, **geospatial intelligence** to deliver **personalized recommendations** and spatial experiences - [Demo](#) | [GitHub](#)