ARYAN MAHESHWARI

Los Angeles, California | (213) 272-9648 | aryan26.03.02@gmail.com | 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, SQL, C++, JavaScript/TypeScript; proficient in relational (PostgreSQL) and NoSQL databases.
Libraries/Frameworks: NumPy, Scikit-learn, TensorFlow, PyTorch, Keras, Matplotlib/Seaborn; backend (FastAPI, Flask, Django)
Infrastructure/Cloud: AWS, GCP, Azure, REST API's, MLOps, Langchain, CUDA, MCP, Ollama, Langgraph, Data Pipelines, git

PROFESSIONAL EXPERIENCE

Convexia (YC'25) San Francisco,California
AI Engineer *July 2025-Present*

- Created end-to-end toxicity evaluation pipeline integrating **6** ML models with MLflow tracking, SHAP-based feature visualization, and confidence/disagreement detection across organ-toxicity modules, ensuring **100%** reproducibility.
- Streamlined infrastructure by modularizing training and inference workflows, implementing structured logging, and reducing CI/CD runtime by **30%** through optimized directory design and automated quick-start setup.

USC Games Los Angeles,California
Machine Learning Engineer *June 2025-September 2025*

- Built hybrid coach selection engine integrating rule-based constraints with AI-driven scoring models, dynamically evaluating **15+ attributes** to improve team matching accuracy by **25%** and reduce lineup imbalance by **40**%.
- Led cross-functional team of **8** interns to design AI-powered sports simulation platform, overseeing model architecture, data pipelines, and deployment to deliver a production-ready system.

USC Autodrive Lab Los Angeles,California
Machine Learning Engineer *June 2025-Present*

- Engineered perception, motion prediction, and planning models for autonomous driving; deployed transformer-based generative AI on NVIDIA CUDA clusters, achieving **4** times faster training throughput.
- Programmed Deep RL algorithms (PPO, SAC) for vehicle navigation, attaining **0.4** m mean positional deviation across **500+** closed-loop test runs.

AGIE AI Mumbai, India
AI Engineer *August 2024-November 2024*

- Developed and deployed Dialogflow-based chatbot with Vertex AI integration, reducing response latency by **23%** (**1.3**s → **1.0**s) and improving engagement across **2** pilot client campaigns.
- Directed a team of **3** interns to deliver Proof of Concept leveraging GPT-**3.5** for semantic similarity scoring and NLP-based retrieval, achieving >**80**% relevance precision for mapping AI research insights to funded startups.

Exicom Technologies Mumbai, India
Machine Learning Engineer *August 2023-December 2023*

- Optimized ML data pipelines for airborne communication systems, reducing latency by **20%** (**250ms → 200ms**) and improving experiment efficiency by **25%** through API integrations with PostgreSQL.
- Tuned feature store queries and indexing strategies, cutting execution time by **35%** under high concurrency, ensuring scalability for training and batch inference.lll

Dawn Digitech Mumbai, India
Machine Learning Engineer *February 2023-May 2023*

- Researched AI-driven interview systems, synthesizing insights from **20+** academic papers and open-source projects; engineered preprocessing pipelines boosting sentiment model accuracy by **15%**.
- Devised and deployed sentiment analysis models with **BERT** and **VADER** to classify polarity, leveraging TensorFlow/PyTorch for NLP workflows and advancing team-wide ML capability by **30%**.

PROJECTS

EduMate.ai (GitHub)

- Engineered AI-powered educational platform with **RAG**-based Q&A, quiz/flashcard generation, and real-time chat, integrating GPT-4, LangChain, Chroma, FastAPI, and Next.js, enabling **10,000** + pages of textbook ingestion, sub-**2**s query latency, and scalable deployment on AWS Fargate..

HieQue (GitHub)

- Developed a scalable multi-level text retrieval framework integrating Gaussian Mixture Models, **GPT-4**-turbo, **BM25**, and semantic search (**SPIDER**), enabling granular content extraction from **300+** page academic textbooks while improving re-ranking precision by **30**% and ensuring low-latency query execution in research-intensive environments.

HistoHelp (GitHub)

- Devised end-to-end histopathology image classification pipeline using **MobileNetV2**, achieving **92%** accuracy on IDC detection; integrated Grad-CAM for interpretability, deployed DCGAN for synthetic data augmentation, and built interactive Streamlit app for real-time predictions with modular, production-ready architecture.