AISHA SARTAJ

Los Angeles, CA ● aishasartaj1@ucla.edu ● 424-483-9810 ● LinkedIn

EDUCATION

Master of Engineering, Data Science (GPA: 3.75/4.00)   **Expected Dec 2025**
University of California, Los Angeles (UCLA)
Relevant Coursework: Large-Scale Machine Learning, Natural Language Processing, Neural Networks and Deep Learning, Big Data Systems, Data and Business Analytics, Management Communication, Product and Project Management

Bachelor of Technology in Computer Science and Engineering (CGPA: 8.71/10.00)   **Jul 2018 - Jul 2022**
Vellore Institute of Technology, Vellore, India (VIT)

TECHNICAL SKILLS

Programming Languages: Python, SQL, R, JavaScript, HTML, CSS
AI/ML Frameworks & Libraries: PyTorch, TensorFlow, Keras, Scikit-learn, Hugging Face Transformers, LangChain, LangGraph, OpenAI API, Pandas, NumPy, Matplotlib, Seaborn, Streamlit, React
MLOps & Model Deployment: MLflow, Kubernetes, Docker, CI/CD for ML models, Model monitoring, A/B testing, TensorBoard, Weights & Biases, DVC (Data Version Control)
Cloud & Infrastructure: AWS (SageMaker, EC2, S3, Glue, Lambda, Bedrock, API Gateway, DynamoDB), Docker, Kubernetes, Apache Spark, PySpark
Databases: PostgreSQL, MySQL, MongoDB, Redis, Snowflake, DynamoDB, Vector databases (Pinecone/Chroma/OpenSearch Serverless)
Development Tools: Git, Jupyter, VS Code, Linux/Unix, Airflow
LLM & GenAI Engineering: LLM fine-tuning & training (LoRA, full fine-tuning), Prompt engineering, Model optimization, Inference serving, Multi-agent systems, Retrieval-Augmented Generation, Guardrails, Vector search & embeddings, API integration & orchestration

WORK EXPERIENCE

ILMAscent (SocratifyMe), AI Engineer   **Mar 2025 – Present**
- Architecting production multi-agent AI system using LangGraph for state-based orchestration of 6 specialized agents with conditional routing, cyclic workflows, and shared memory management, replacing rigid state machines with dynamic graph-based execution for adaptive educational feedback analysis and real-time pedagogical assessment generation.
- Engineering advanced RAG pipeline implementing 8 retrieval patterns (simple semantic search, multi-query expansion, hybrid dense-sparse retrieval, self-RAG with confidence thresholds, agentic retrieval with tool use, corrective RAG with re-ranking) using OpenSearch Serverless vector store, Amazon Bedrock LLMs (Nova Pro for cost optimization, Claude Sonnet 4.5 for accuracy), and AWS Textract for multimodal document intelligence with OCR confidence scoring and mathematical notation extraction.
- Building production MLOps infrastructure with automated evaluation frameworks (no human-in-loop), LangSmith observability for agent trace analysis, containerized Lambda deployment via SAM/Docker, semantic caching (60% cost reduction), intelligent model routing, CI/CD via AWS CodePipeline, distributed tracing (X-Ray/CloudWatch), prompt versioning with A/B testing, and security controls (IAM/KMS/Bedrock Guardrails) ensuring FERPA compliance and bias detection across demographic groups.

Institute of Applied Artificial Intelligence and Robotics (Remote, US), Applied AI Research Intern   **Jun 2025 – Sep 2025**
- Delivered a color grading pipeline for emerald gemstones by curating and calibrating a 1,000+ image dataset with white-balance normalization, glare masking, and continuous LAB hue features, achieving 97% Random Forest CV accuracy and 95%+ SVM accuracy across bluish-green, pure green, and yellowish-green classes.
- Built an automated clarity assessment system by manually annotating 37 emerald images and later expanding to 1,000+ samples, applying CLAHE preprocessing, edge/texture feature extraction, and Random Forest/SVM modeling, reaching ~80% CV accuracy despite challenges of data imbalance and noisy labels.
- Developed an end-to-end AI-powered gemstone grading pipeline combining classical CV (GrabCut, CLAHE, morphological cleaning) with deep learning (ResNet50 feature extraction, GAN-based augmentation, ensemble RF/SVM) and explainable AI (Grad-CAM), enabling scalable, expert-level gemstone evaluation with reproducible workflows.

Leadoff.ai (Remote, US), Data Science Capstone Intern   **Jun 2025 – Aug 2025**
- Integrated Hume.ai's multimodal sentiment analysis into behavioral reasoning pipeline by engineering time-range aggregation functions that compute primary emotions and top-5 composite scores across 4 modalities (speech prosody, facial, vocal burst, language), injecting emotional context into timestamped buyer behavior instances for sales coaching insights.
- Transformed text-based behavioral explanations into granular timestamp-enriched reasoning objects with start/end timestamps and speaker attribution, achieving 100% speaker classification accuracy (19/19 instances) through fuzzy text matching (rapidfuzz) and time-based alignment, enabling emotion-to-behavior correlation across sales conversations.
- Evaluated 6 multimodal sentiment platforms (Hume.ai, IBM Watson, Azure AI Video Indexer) across text/audio/video capabilities, recommending Hume.ai for production deployment and implementing 6-collection MongoDB schema (HumeProsody, HumeFace, HumeLanguage, HumeVocalBurst, HumeSpeakerComposite, HumeJobs) for emotional intelligence storage.

Fractal Analytics (Mumbai, India), AI Software Engineer [Enterprise Solution]   **Aug 2023 – Sep 2024**
- Prototyped a Generative AI-powered text-to-SQL solution using LangChain and LLMs, replacing the homepage search with NLP-based queries to improve data accessibility and user experience.
- Built a React-based chatbot using Avalok.ai APIs to process natural language queries and render dynamic tables, charts, and code; integrated Power BI dashboards for insights—boosting data engagement by 40% and presented to the CAIO and CTO.
- Designed GenAI prompts with product and architecture leads to auto-generate clear tables, KPIs, and definitions—cutting data interpretation time by 30% and improving user satisfaction by 25%.

Fractal Analytics (Mumbai, India), Data Engineer   **Sep 2022 – Jul 2023**
- Built and optimized robust ETL/ELT pipelines using AWS (S3, Glue, Lambda, Airflow), SnapLogic, and Snowflake—boosting pipeline reliability and scaling client project revenue from $75K to nearly $500K through performance tuning, error handling and real-time monitoring with CloudWatch.
- Strengthened the client's architecture by developing a reconciliation framework for BOTREE and SAP systems—automating discrepancy resolution and later standardized across systems—and improved Glue logging and SnapLogic API integration to enhance data sharing and system reliability.
- **Achieved AWS Solutions Architect (SAA-C03) certification** and built pipelines using AWS (S3, RDS, Glue, Lambda, DynamoDB, EMR, Redshift) during training, enhancing data processing and integration; earned a 93% onboarding score.

Fractal Analytics (Mumbai, India), Data Science Apprentice   **May 2022 – Jul 2022**
- Engagement Prediction System: Developed a video engagement prediction system leveraging machine learning models like XGBoost, linear regression, and random forests. Selected XGBoost for its superior R^2 and RMSE scores to deliver an efficient solution in June 2022.
- Technical Skills: Explored and worked on problems using Excel, SQL, Data Exploration, Statistical Inference, Univariate and Bivariate analysis using Python, Power BI- Power Query, Machine learning, Data Engineering, Spark, Spark streaming, MongoDB, and Azure.

DS/ML Projects

AI Ingredient Analyzer App   **July 2025 – Present**
- Architecting a multi-layered AI system with FastMCP caching, Vectorize.io RAG vector database for semantic ingredient matching, Tavily API for real-time EWG data scraping, and fallback LLM orchestration to ensure 99%+ ingredient coverage with sub-2s response times through intelligent cache-first retrieval strategy.
- Building a multi-agent pipeline using Python orchestrator that handles ingredient extraction, safety analysis, and interaction detection with structured output validation, guardrails for input sanitization, and personalized recommendations based on user skin profiles stored in vector embeddings.
- Implementing Streamlit frontend with interactive ingredient modals, safety rating visualizations following EWG standards, and a comparative analysis feature that uses semantic similarity scoring to detect conflicts/synergies between product formulations for layering compatibility checks.

MoodFlow [Github]   **Oct 2025**
- Built emotion-aware AI scheduler using AWS Bedrock (Claude 3.5), Knowledge Base RAG, and OpenSearch Serverless that adapts task ordering and time blocks based on user mood, supporting 7 emotional states with research-backed scheduling strategies.
- Implemented serverless architecture with Lambda orchestration, DynamoDB persistence, API Gateway REST endpoints, and Bedrock Guardrails, achieving <20s response times and handling cross-date rescheduling with conversational interface.
- Designed RAG system using 3-document knowledge base with vector similarity search to retrieve evidence-based planning strategies, ensuring AI recommendations stay grounded in psychological research rather than improvisation.

AllyIn Compass – AI-Powered Enterprise Assistant [GitHub]   **May 2025 – Jun 2025**
- Built a multi-agent assistant using LangChain, Streamlit, and Neo4j to answer enterprise queries across SQL, vector, and graph data.
- Simulated fine-tuning with LoRA (PEFT) using synthetic feedback for domain-specific tasks.
- Designed a real-time observability dashboard to monitor tool usage, latency, and query performance.

EMG Keystroke Prediction: Conformer vs. BiLSTM, UCLA   **Jan 2025 – Mar 2025**
- Performed comparative analysis of Conformer and BiLSTM architectures for EMG-based keystroke prediction using the emg2qwerty dataset.
- Achieved 23.3% Character Error Rate(CER) with Conformer, compared to 14.93% with CNN+BiLSTM and 30% from the TDS baseline.
- Identified optimal settings (e.g., 1500Hz sampling, 12 channels), and found that BiLSTM better captures EMG timing dynamics than Conformer's self-attention.

Publication

Pioneering Image Analysis with Hybrid CNN and GAN for Enhanced Visual Perception, IDSCS 2023 [Link]