Moksha Shah

408-420-5253 | San Francisco, CA | mokshashah1112@gmail.com | linkedin.com/in/moksha0111 | github.com/mokshashah0111

Summary

Aspiring Software Engineer specializing in AI/ ML Infrastructure and Backend development. A team player with excellent communication skills, thriving in collaborative environments, and dedicated to building scalable, user-centric applications.

Projects

Personality Classification | Azure ML Studio,C++, K-Means Clustering, Decision Tree Classifier

- Developed an end-to-end ML pipeline in **C++** and **Azure ML Studio** with **MLTable** ingestion, feature scaling, label handling, and train/test split; applied **K-Means (k=3)** and **Multiclass Decision Forest** on a 20K×30 dataset.

Cold Email Generator | Python, LLaMa, Langchain, Streamlit, ChromaDB, GROQ

- Built an email-generation pipeline with **Llama** on **Groq** using **LangChain**, and **Streamlit**, reducing outreach by **80%** and achieving **90%** job-to-portfolio match accuracy using **ChromaDB** vector search with **cosine-similarity** retrieval.

AI Chatbot | Python, LLaMa, Gradio, AWS, Flask

- Built a **Flask**-based conversational AI with **Llama**, fine-tuned via **Keras** and **Gradio** UI achieving **85%** accuracy; deployed with Docker on **AWS EC2** for scalable, high-availability performance.

Facial Recognition | Python, CNN, TensorFlow, PyTorch, OpenCV

- Built an emotion-recognition pipeline achieving **77%** accuracy on 20K+ images across 7 classes, and cut latency by **30%** through cross-validation, augmentation, and **TensorRT**/**OpenCV** optimization.
- Applied **5-fold** cross-validation, edge case testing, and class-balancing augmentation, cut inference latency by **30%** via **OpenCV** threading and **TensorRT**.

Music Artist Recommender System | Python, Collaborative Filtering, Natural Language Processing (NLP), Implicit

- Achieved **94%** precision on top-N artist recommendations by combining **Collaborative Filtering** with **NLP**-based **tokenization** of artist metadata. .
- Accelerated training by **40%** using the **Implicit** library on **Apache Spark**; evaluated with **MAP@10** and **RMSE** and tuned hyperparameters to improve cold-start support.

Experience

Software Engineer | **Inexture Solutions LLP. (India)** **Jan 2024 – July 2024**

- Reduced the latency by **35%** for **Django** services using async request handling and modular refactoring; developed production-ready backend services using **REST APIs** across microservice architecture, improving scalability by **15%**.
- Boosted **MySQL** execution by **15%** through advanced query tuning, indexing strategies, and partitioning of large datasets.
- Provisioned and maintained **AWS infrastructure (EC2, S3)** in real-time ML services; automated **CI/ CD** with **AWS CodePipeline** and **GitHub Actions**, enabling fast rollbacks, increasing release cycles by **40%**.
- Collaborated with cross-functional teams to manage system-level metrics and updates for **backend** performance.

Research Assistant | **California State University- East Bay (USA)** **Aug 2025 – Present**

- Training the **OpenVLA** model on the **Dofbot Pro** robot to execute real-world actions based on human commands.
- Using **Phosphobot** and **Python** for robot control, data logging, and analysing model predictions with physical movement.
- Designing and refining the robot's **vision** and control systems to improve response accuracy and action execution.
- Iterating on model training and system performance to make tasks more precise and reliable.

Technical Skills

Languages: Python, C/C++, Go, Java

Frameworks and Tools: ReactJS, Django, Flask, Streamlit, Gradio, Phosphobot

Machine Learning & AI: PyTorch, TensorFlow, Scikit-learn, HuggingFace, LangChain, LLaMA, VLM (OpenVLA), Generative AI (OpenAI, Claude, GPT-4o, Ollama), Azure ML Studio, Transformers, NLP, Computer Vision

Deployment: FastAPI, Rest API, Vertex AI, Docker, Kubernetes, Terraform, Spark, Hadoop

ETL & Data Tools : SQL, Airflow, BigQuery, MLFlow, Flyte, OAuth, VectorDB (FAISS, ChromaDB), MySQL, PostgreSQL

Cloud & DevOps: AWS (EC2, S3, Lambda), Azure, GCP

Other: Git/ GitHub, Compiler Design, Distributed Systems, Data Structures and Algorithms, Object-Oriented Programming

Education

California State University, East Bay | Master of Computer Science (GPA: 3.9) Expected: May 2026

Gujarat Technological University, India | Bachelor of Computer & Electronics Engineering (GPA: 4.0) 2020 - 2024