

Shubham Chaudhari

Los Angeles, CA | shubhamc@usc.edu | [linkedin.com/in/shubham-ch](https://www.linkedin.com/in/shubham-ch) | github.com/Esoteriikos | [My Portfolio](#) | 213-275-7114

SUMMARY

Graduate student at USC with 3.5+ years of experience in MLOps, LLMops, and GenAI. Designed full-stack AI pipelines with data engineering, orchestration, observability, and data flows. Built tools composing code generation, knowledge retrieval, and decision-making across teams. Recognized with multiple awards for driving business innovation on real-world challenges

SKILLS

- **Languages:** Python | **Core:** AI/ML Systems Design, MLOps/LLMops, Computer Vision, GenAI, Data Science, AI Ethics, Transformers, XGBoost, PyTorch, TensorFlow, Scikit-learn, OpenCV, NumPy, Pandas
- **LLM & Generative AI:** RAG, AI Agents, (LangChain, LlamaIndex, LlamaParse), RAG Evals (Ragas), LangGraph, Fine-Tuning, LoRA, QLoRA, Prompt Engineering, Multi-modality, GANs, Azure Vision Studio, VectorDB, Guardrails
- **DevOps:** Docker, Git, GitHub Actions, DVC, FastAPI, Flask, Streamlit, MLflow, Postman, Linux
- **Databases:** PostgreSQL, MongoDB, Redis, MySQL | **Orchestration & ETL:** Kafka, Airbyte, n8n, Airflow
- **Monitoring & BI:** Metabase, Grafana, Plotly, Seaborn | **Optimization:** ONNX, TensorRT, OpenVINO

EDUCATION

UNIVERSITY OF SOUTHERN CALIFORNIA

Master of Science in Computer Science - Data Science Specialization (GPA 3.65/4.00)

Los Angeles, CA

January 2025 - December 2026

UNIVERSITY OF MUMBAI

Bachelor of Engineering, Computer Engineering (CGPA 9.14/10)

Mumbai, India

July 2017 - June 2021

EXPERIENCE

HERE TECHNOLOGIES, Technology Innovation Labs

Mumbai, India

AI ML Engineer II

August 2024 - December 2024

- Engineered context-aware AI coding assistant by fine-tuning CodeQwen and implementing a LangChain RAG pipeline querying proprietary Java SDKs from a vector database, all managed within a robust LLMops framework, leading to a 30% rise in productivity
- Constructed data ingestion and ETL pipeline for RAG, parsing multiple file formats with a hybrid approach of LlamaParse and custom scripts improving context-aware chunking to preserve semantic integrity, diminishing LLM token overhead by 40%
- Implemented a cross-encoder re-ranker within a hybrid retrieval system with prompt fine tuning and guardrails for response quality achieving 40% lift in contextual accuracy over baseline vector search and providing cleaner context to LLM generator model

TATA CONSULTANCY SERVICES, Retail Strategic Initiatives Lab

Mumbai, India

AI ML Engineer

July 2021 - July 2024

- Led development of multi-camera vision pipeline integrating OpenPose, OpenCV with data from IoT load sensors using Pub/Sub messaging system to track end-to-end customer journey in autonomous store, decreasing checkout times by over 80%
- Developed two targeted solutions utilizing Azure Vision Studio: Custom shelf-monitoring model cutting stockout and spoilage delays by 65%, and CCTV video summarization and text querying service fetching incident reducing incident review time by 70%
- Devised generative AI pipeline for automated cataloging, enabling 2D virtual try-ons via Super-Resolution CNNs and GANs cutting production time from weeks to hours, while prototyping 3D reconstruction (NeRFs/Splatting) creating assets from 2D images
- Optimized models for edge deployment on NVIDIA Jetson devices, using ONNX, TensorRT and model quantization boosting real-time item recognition achieving 7x faster inference
- Spearheaded collaborative efforts 'TechFM' knowledge-sharing program, training 400+ colleagues and 1000+ Bachelor's students in practical AI/ML concepts, fostering a culture of technical excellence and continuous learning

TATA CONSULTANCY SERVICES

Mumbai, India

Data Scientist Intern

February 2021 - May 2021

- Led team of 5 and owned end-to-end Materiovigilance system accelerating adverse event detection by 60%; integrating MongoDB, Lasso/XGBoost predictive model with a real-time Streamlit dashboard providing actionable insights for faster clinical response

PROJECTS

- Crafted real-time data pipeline using Airbyte, Kafka, and PostgreSQL deliver analytics and business insights on a Metabase dashboard
- Monocular Depth Estimation: Built U-net pipeline extracting depth information with Object detection, and 3D scene reconstruction

PUBLICATIONS

- Published Whitepaper: 'A Hybrid approach towards Signal Management in PV' ([link](#))
- 'A Synopsis of Monocular Depth Estimation' (Presented at ICDSMLA, Published by Springer) ([link](#))