# Shubham Chaudhari

Los Angeles, CA | shubhamc@usc.edu | 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 SDLC, 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 (Primary), Javascript, React, HTML, CSS, R | **Core**: AI/ML Systems Design, MLOps/LLMOps, Computer Vision, GenAI, NLP, 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, MCP
- **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, Prometheus | **Optimization**: ONNX, TensorRT, OpenVINO

## EDUCATION

**UNIVERSITY OF SOUTHERN CALIFORNIA** — **Los Angeles, CA**
**Master of Science in Computer Science - Data Science Specialization ( GPA 3.65/4.00)** — **January 2025 - Present**

**UNIVERSITY OF MUMBAI** — **Mumbai, India**
**Bachelor of Engineering, Computer Engineering ( CGPA 9.14/10 )** — **July 2017 - June 2021**

## EXPERIENCE

**USC FPM - MIS | Backend System Engineer [Student] | Los Angeles, CA, USA** — **June 2025 - Present**
- Restructuring and optimization of the MIS system, enhancing scalability and fault tolerance for backend API workflows in production
- Collaborated on MIS data pipelines, enabling real-time analytics and supporting expansion of operational dashboards

**SIGMA HEALTH SENSE | Full Stack AI Engineer [Volunteer] | Los Angeles, CA, USA** — **June 2025 - August 2025**
- Implemented TTS/STT solutions for custom AI agents and chatbots, integrating Twilio to enhance business workflow automation
- Built modular AI pipelines to streamline voice-data processing, leveraging Gemini to extract user intent and trigger dynamic actions like querying health databases to assist patients in real time

**HERE TECHNOLOGIES | AI ML Engineer II | Mumbai, IN** — **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 | AI ML Engineer | Mumbai, IN** — **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

**TATA CONSULTANCY SERVICES | Data Scientist Intern | Mumbai, IN** — **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
- Real-time anomaly detection pipeline for transactions using Python, Kafka, and FastAPI, leveraging statistical feature engineering, scikit-learn models, CI/CD and MLOps tools (MLflow, Prometheus, Grafana) for end-to-end monitoring and model management
- Conversational command center that streamlines Git, Docker, and project workflows via a unified MCP, LLM-powered interface

## PUBLICATIONS

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