Soham Chaudhari

+91 7499787431 | sohamrc08@gmail.com | LinkedIn | GitHub | Portfolio

SUMMARY

AI Engineer with production experience in building scalable Agentic AI systems and RAG architectures. Proficient in PyTorch, FastAPI, and React, with a track record of deploying high-performance ML models (95% accuracy) and optimizing API latency by 40%.

EDUCATION

B.E. Computer Engineering | *Nov 2022 - Expected 2026*

Vivekanand Education Society's Institute of Technology | **CGPA: 8.11**

WORK EXPERIENCE

SDE Intern, Mumbai | SR Counselling | *Dec 2024 - Sep 2025*

- Implemented and deployed ML recommendation system: 95% accuracy on 50K+ products
- Engineered high-concurrency REST APIs using Node.js/Express, reducing latency by 40% (500ms to 300ms) for 10k+ concurrent users & 99.8% uptime.
- Lead the development of an automated visa training system integrating TTS/STT pipelines, reducing manual mock interview time.

AI Developer, Remote | UnLawC | *Sep 2025 - Present*

- Designed and Deployed end-to-end Legal Ops AI agents, reducing manual document review time by almost 50%.
- Optimized LLM performance using advanced prompt engineering, improving contextual accuracy by 31% across legal query evaluations.
- Architected RAG pipelines with LangChain and FAISS, increasing retrieval accuracy by 60% and eliminating hallucinations via self-correction loops

PROJECTS

VISION AI - AI-based Image and Video Upscaling | GitHub(url) | *Oct 2024 - Oct 2024*

Technologies: ReactJS, Flask, PyTorch, OpenCV, ESRGAN, Computer Vision, Deep Learning, GANs (Generative Adversarial Networks), Transfer Learning, Model Evaluation

- Pioneered an AI-based upscaling solution using ESRGAN architecture and transfer learning, achieving 4x resolution enhancement for images and CCTV footage with 92% quality improvement
- Applied **GANs** using PyTorch on the DIV2k dataset, processing images with a 92% quality improvement rate. Performed error analysis using confusion matrices, **ROC curves**, and **hyperparameter tuning (GridSearchCV, Optuna).**
- **Deployed** the backend on **Render** and the frontend on **Vercel**.

MOSAIC - AI-Powered Multimodal Video Analysis Platform | GitHub(url) | *Dec 2025- Jan 2026*

Technologies: FastAPI, React, FastMCP, FFmpeg, Groq Whisper, CLIP, FAISS, ChromaDB, LangChain, Vector Embeddings, Semantic Search, Multimodal AI, ReAct Agents, asyncio

- Architected a multimodal video search engine using **CLIP** and **FAISS**, achieving **99.9% recall** for natural language video queries and containerized using **Docker**.
- Built Agentic workflows using **LangGraph**, enabling autonomous video segmentation and timestamp extraction with 90% precision.
- Optimised video processing pipeline using **asyncio** and **FFmpeg**, supporting 5+ simultaneous streams with **less than 500ms latency** on consumer hardware.

AI Bootstrap - Generative AI Project Scaffolding | PyPi(url) | *July 2025- October 2025*

Technologies: Python, Typer, Copier, Docker, LLMs, CLI Development, Jinja Templating, and LangChain

- Engineered a production-ready CLI tool (published on PyPi) that scaffolds AI/ML projects in <30 seconds, adopting Clean Architecture principles.
- Automates environment setup using Docker and Poetry, currently used by developers to reduce project initialization time by 90%.
- Published on PyPi and wrote comprehensive documentation for better collaboration and contributions.

SKILLS

Programming Languages: Python, JavaScript, Java, SQL

AI & LLMs: LangChain, LangGraph, Retrieval-Augmented Generation, Prompt Engineering, Transformers, CLIP, Multimodal AI

Machine Learning & Deep Learning: PyTorch, TensorFlow, scikit-learn, Neural Networks, Computer Vision, NLP

Databases & Semantic Search: FAISS, ChromaDB, Pinecone, MongoDB, PostgreSQL

MLOps & Deployment: Docker, MLflow, CI/CD Pipelines, Git, GitHub

Web & APIs: FastAPI, Flask, Node.js, Express.js, React.js, Next.js