

SOHAM CHAUDHARI

+91 7499787431 — Mumbai, Maharashtra, India — sohamrc08@gmail.com — [LinkedIn](#) — [GitHub](#)

SUMMARY

AI Developer with hands-on experience building Agentic AI systems, RAG architectures, and multimodal AI pipelines. Proven impact optimizing API latency and reducing manual workflows. Seeking AI/ML Engineering roles in Generative AI, NLP, and Computer Vision.

EXPERIENCE

AI Developer Sep 2025 – Present
UnLawC *Remote*

- Designed and deployed end-to-end Legal Ops AI agents, reducing manual document review time by 50%
- Optimized LLM performance using advanced prompt engineering, improving contextual accuracy by 31%
- Architected RAG pipelines with Mistral, LangChain and Pinecone, increasing retrieval accuracy by 60% while eliminating hallucinations

SDE Intern Dec 2024 – Sep 2025
SR Counselling *Mumbai*

- Developed ML recommendation system achieving 95% accuracy across 50K+ products using collaborative filtering
- Engineered high-concurrency REST APIs with Node.js/Express, reducing latency by 40% for 10K+ concurrent users with 99.8% uptime
- Built automated visa training platform integrating TTS/STT pipelines, streamlining mock interview processes

EDUCATION

B.E. Computer Engineering Nov 2022 – Expected 2026
Vivekanand Education Society's Institute of Technology — CGPA: 8.11

PROJECTS

MOSAIC – AI-Powered Multimodal Video Analysis Platform [GitHub](#)

- Architected multimodal video search engine using CLIP and FAISS, achieving 99.9% recall for natural language queries
- Built agentic workflows with LangGraph enabling autonomous video segmentation with 90% precision
- Enhanced processing pipeline using asyncio and FFmpeg, supporting 5+ concurrent streams with sub-500ms latency

VISION AI – Image and Video Upscaling [GitHub](#)

- Developed ESRGAN-based super-resolution system achieving 4x resolution enhancement with 92% perceptual quality improvement
- Applied GAN training using PyTorch on DIV2K dataset with hyperparameter optimization via Optuna
- Deployed scalable backend on Render and frontend on Vercel for production use

AI Bootstrap – Generative AI Project Scaffolding Tool [PyPI](#)

- Engineered production-ready CLI tool published on PyPI that scaffolds AI/ML projects in under 30 seconds using Clean Architecture
- Automated Docker and Poetry-based environment setup, reducing project initialization time by 90%

SKILLS

Programming Languages	Python, JavaScript, Java, SQL
AI & LLMs	LangChain, LangGraph, RAG, Prompt Engineering, Transformers, CLIP
ML & Deep Learning	PyTorch, TensorFlow, scikit-learn, Computer Vision, NLP
Databases & Search	FAISS, ChromaDB, Pinecone, MongoDB, PostgreSQL
MLOps & Deployment	Docker, MLflow, CI/CD, Git, GitHub
Web & APIs	FastAPI, Flask, Node.js, Express.js, React.js, Next.js

ACHIEVEMENTS

- Runner-up at Medino X Advitiya'25 Hackathon, IIT Ropar (2025), among 200+ participants for healthcare AI solution
- Top 10 finalist at Hack2Future 2024, IIIT Dharwad, among 400+ participants for Vision AI solution