

Armaan Irani

+91-9921494567 | armaanirani@gmail.com | www.github.com/armaanirani
www.linkedin.com/in/armaanirani
Pune, Maharashtra

Education & Certificates

- M.Sc. in Data Science & AI from University of Liverpool, Liverpool Graduated in 2024
- B.Sc. in Mathematics from Pune University, Pune Graduated in 2023

Work History

Software Engineer Intern, Script Lanes, Pune

May 2026 to Current

- Shipped a medical e-learning platform, streamlining development using Claude Code to build the TypeScript, Node.js, and Express backend with PostgreSQL on AWS RDS that turns digital and handwritten notes into organized, stored study material.
- Built a Python and FastAPI microservice that applies prompt engineering to guide a VLM through note transcription and classification in a single structured pass, backed by Zod schema validation and guardrails that keep incorrect medical content from reaching users.
- Designed a RAG system for the LLM-powered chatbot students ask about their syllabus, using a lightweight classifier model to route each question to the relevant curriculum folders before a larger model answers, grounded in the complete, human-authored course material.
- Routed requests across Gemini and OpenAI GPT models through OpenRouter based on task complexity to keep inference costs low, iterating directly with medical student testers and an LLM-as-a-judge evaluation process to keep responses accurate before shipping.
- Deployed the platform's services on AWS using EC2 behind an Application Load Balancer, with S3 file storage under per-user access permissions and short-lived signed URLs, keeping the app reliable in production and every user's uploaded notes private.

Machine Learning Engineer, University of Liverpool, Remote

June 2024 to Current

- Owned an end-to-end machine learning pipeline in Python, from raw detector readings to a shipped, evaluated model that determines radiation direction from a single sensor, locating a source without multiple detectors.
- Partnered with nuclear physicists in a cross-functional collaboration to build leakage-free training and test datasets and robustness-test real-world field data requirements.
- Built an automated, Git-versioned evaluation harness in Python with scikit-learn, training and scoring model and feature-extraction combinations in parallel on a compute cluster under standardized reporting.

- Tuned TensorFlow and PyTorch neural networks with Bayesian hyperparameter optimization across an open-ended, poorly-specified problem space, then shipped a simpler classical baseline that outperformed them, optimizing inference costs.

AI Engineer Intern, Nextoar, Bangalore

June 2026 to Current

- Working directly with the founder and CTO at an early-stage industrial AI startup, building AI agents, computer vision apps, and LLM-powered tools while contributing to business and strategic decisions as the company grows.
- Improving an existing agentic outreach tool by adding custom tool calling and connecting to MCP servers using Python and OpenAI SDK, researching and drafting personalized emails for prospects actively facing problems in Nextoar's target industries, keeping leads targeted and sales-ready instead of generic cold outreach.
- Fine-tuning the computer vision models behind an industrial floor assistant for its detection needs, and added hybrid retrieval to its RAG system so troubleshooting answers surface exact part numbers and error codes.
- Rebuilding Nextoar's public website using TypeScript, 11ty.js, and Tailwind CSS, replacing a fragmented WordPress site with one consistent build the whole company can maintain.
- Integrating Langfuse tracing and model routing across LLM providers throughout the building lifecycle, catching failures early and keeping inference costs down as usage scales.

Projects

Agentic AI Compliance Engine

Feb 2026 to March 2026

- Built an AI agent with LangGraph and FastAPI that reviews financial videos for compliance issues, freeing compliance teams from watching footage manually.
- Grounded the agent's answers in real transcripts and documents using hybrid RAG with Azure AI Search and reranking, cutting down hallucinated or inaccurate rules.
- Tracked and debugged agent runs with LangSmith observability, and validated outputs against structured schemas using an LLM-as-a-judge layer, providing visibility to catch failures before they reach production.

GPT-2 Small from Scratch

June 2025 to July 2025

- Built GPT-2 Small from scratch in PyTorch to understand transformer architecture and attention mechanisms first-hand, beyond just calling existing LLMs through APIs.
- Used Top-k sampling and temperature scaling to reduce repetitive, made-up text, applying the same generative AI techniques production systems rely on.