Harsh Virani

AI & ML Engineer | Full-Stack Prototyping | RAG & Agentic Systems

Hof, Germany | +49 15510862272 | harshvirani.91@gmail.com
LinkedIn | GitHub

M.Sc. AI & Robotics | Production-oriented AI from RAG & agents to computer vision | Full-stack delivery (React, Next.js, FastAPI) | Dean's List 2025

Technical Skills

GenAI & Agents:	LangChain, LangGraph, RAG, LangSmith, Prompt Engineering, Groq, Google Gemini, Llama 3.3, MCP
ML & Deep Learning:	PyTorch, TensorFlow, Scikit-learn, XGBoost, Transfer Learning, CNNs (EfficientNet, ResNet, DenseNet), Model Benchmarking
Full-Stack & APIs:	Python, FastAPI, React, Next.js, TypeScript, Node.js, REST, Pydantic, PostgreSQL
MLOps & Cloud:	Docker, AWS (S3, CloudFront, CodePipeline), CI/CD, FAISS, Qdrant, Vercel, Render
Data & Evaluation:	ETL Pipelines, Cross-Validation, AUC/Precision/Recall/F1, Pandas, NumPy, Hugging Face

Education

Master of Science in AI & Robotics — **March 2025 – Present**
Hof University of Applied Sciences — *Hof, Germany*
GPA: 1.8/5.0 | **Dean's List 2025** | Expected Graduation: March 2027 | Key Modules: Generative AI, Machine Learning, Deep Learning, Data Mining, Advanced AI Applications, Research Methodology & Academic Writing

Bachelor of Science in Computer Science — **August 2021 – March 2024**
Veer Narmad South Gujarat University — *Gujarat, India*
CGPA: 9.01/10.0 | Key Modules: Data Structures & Algorithms, Software Engineering, Database Systems, Web & Mobile Development

Projects — GenAI, Computer Vision & Full-Stack

Agentic AI Customer Support System — **GitHub | Live Demo**
Multi-Agent RAG System — *2026*

- Built production GenAI system with LangGraph multi-agent orchestration (classifier + 3 domain agents), FastAPI backend, and Qdrant RAG pipeline; LangSmith tracing for 100% agent observability across latency, routing accuracy, and retrieval scores
- Benchmarked embedding approaches (local 384d vs. API 768d Gemini), reducing deployment size by 93% (2GB $\rightarrow$ 130MB) while improving retrieval quality and maintaining $<$3s end-to-end response latency
- Engineered category-aware query enhancement and top-K semantic retrieval over 97 knowledge-base chunks with source citations for transparent GenAI responses
- **Tech Stack:** Python, LangGraph, LangChain, FastAPI, Qdrant, LangSmith, Groq, Google Gemini, React, Next.js

AI-Powered Thoracic Disease Detection — **GitHub**
Deep Learning & Medical Imaging — *2025*

- Benchmarked 4 CNN architectures on 112K NIH chest X-rays; EfficientNet-B0 (4M params) achieved 83.6% AUC with 2$\times$ parameter efficiency vs. DenseNet-121 baseline
- Raised multi-label recall from 13.7% to 83.2% via class-weighted loss and per-disease threshold tuning across 14 thoracic conditions with patient-level train/val/test splits
- Built automated evaluation (per-class AUC, precision, recall, F1) and OpenCV/PyTorch pipeline with 15ms GPU inference suitable for screening workflows
- **Tech Stack:** Python, PyTorch, OpenCV, Scikit-learn, Transfer Learning

YouTube Video Q&A with RAG Architecture — **GitHub**
RAG Pipeline & Retrieval Optimization — *2025*

- Developed RAG ingestion for YouTube transcripts: RecursiveCharacterTextSplitter (1000/200 overlap), HuggingFace embeddings (768d), FAISS indexing for sub-second retrieval over 10K+ chunks
- Implemented LangChain RunnableParallel workflows, reducing pipeline latency by 45% through concurrent retrieval and query formatting
- Shipped Streamlit prototype with real-time vector store creation, bilingual UI, and session-aware context for rapid retrieval/prompt iteration
- **Tech Stack:** Python, LangChain, FAISS, HuggingFace, Streamlit

SmartDiet AI – Personalized Diet Recommendation
ML + LLM Health Platform *Jan – Apr 2025*

- Benchmarked 3 classifiers (Logistic Regression, Random Forest, XGBoost) on 1,000+ records via 5-fold CV; XGBoost reached 100% accuracy with automated 12+ metric reporting in $<$5s
- Integrated Groq Llama 3.3 70B via LangChain with constraint validation achieving 95% dietary adherence across 20 user profiles
- Delivered Streamlit app with auth, profiles, ML predictions, and LLM-generated weekly meal plans (team capstone, Hof University)
- **Tech Stack:** Python, Scikit-learn, XGBoost, LangChain, Groq, Streamlit, Pydantic

Professional Experience

Software Engineer **July 2024 – March 2025**
TechStaunch Solutions PVT LTD *On-site, India*
GoShimmy E-Commerce Platform:

- Delivered full-stack features for production e-commerce serving 10K+ daily users and 50K+ daily transactions; integrated ASP.NET APIs and SignalR real-time updates
- Architected AWS CI/CD (S3, CloudFront, CodePipeline), cutting deployment time by 87% (8h $\rightarrow$ 50min) across production and beta environments
- Improved page load performance by 40% through profiling, refactoring, and frontend optimization

Research Experience

- **The Evolving Ecosystem of Large Language Models** — comparative analysis of decoder-only, encoder-only, and encoder-decoder architectures across NLU/NLG; prompt strategies, hallucination mitigation, and Code LLMs
- **AI-Augmented Low-Code/No-Code Platforms** — LLM-based agents bridging software engineers and citizen developers; task-technology fit and workflow integration

Academic Projects

- Implemented Auto-Encoding Variational Bayes (AEVB) from research paper to practical prototype as university coursework, bridging theoretical concepts with hands-on development
- Assisted on Heuristical 2D Cloth Warping under faculty guidance, contributing to simulation modeling and algorithmic development

Certifications

- Introduction to Model Context Protocol – Anthropic, March 2026
- Python 101 for Data Science – IBM (Cognitive Class), April 2025

Awards & Achievements

- **Dean's List** – Hof University of Applied Sciences, 2025

Languages

English:	C1 (Business Fluent)
German:	CEFR B1+, actively improving
Hindi:	Native
Gujarati:	Native