HARSH TOMAR

AI/ML Engineer

+91-8817969476 • tomarharsh28303@gmail.com • LinkedIn • GitHub • Portfolio

Professional Summary

Versatile **AI/ML Engineer** with expertise in **Machine Learning**, **Deep Learning**, **Computer Vision**, **MLOps**, **Large Language Models (LLMs)**, and **Generative AI**. Built production ML systems achieving **95% detection accuracy** in real-time video analytics and **88.52% PR-AUC** in fraud detection using ensemble models (XGBoost, LightGBM). Developed agentic AI systems with **94% routing accuracy** using LangGraph. Strong in PyTorch, TensorFlow, LangChain, NLP, end-to-end ML pipelines, and AWS cloud deployments. **40+ GitHub stars** across 47 repositories; **10+ PRs merged** in open-source projects.

Professional Experience

AI Intern **May 2025 – Jul 2025**

i3 Digital Health *Remote*

- **Architected research profiling system** auto-generating researcher profiles by aggregating data from **PubMed, ResearchGate, Google Scholar APIs**
- **Built RAG-powered search agents** using LangChain + vector embeddings for contextual recommendations and research collaborator identification
- Collaborated with healthcare professionals to refine NLP search algorithms; deployed with **FastAPI + Docker**
- Implemented web scraping and **multi-source API integration** for comprehensive research data collection

Community Contributor **Jan 2023 – Present**

CNCF & Google Developer Groups *Remote*

- Presented **2 tech talks** on AI/ML; mentored **10+ developers**; **10+ PRs merged** in open-source (llm-council, kubrick-ai, multimodal-agents)

Projects

Tennis Vision – Real-time Sports Analytics **2025**

PyTorch, YOLOv8, OpenCV, ResNet-50 | Demo | GitHub | **23 Stars**

- Developed real-time tennis analysis: **95% player detection**, **88% ball tracking** using YOLOv8 at **30 FPS**
- ResNet-50 court keypoints (**91.5% accuracy**, 14 landmarks); shot classification for **12 stroke types** (**91%**)
- Optimized memory by **94%** via ROI processing; mini-court visualization with player speed tracking

Decifra – MLOps Fraud Detection Pipeline **2025**

ZenML, MLflow, DVC, BentoML, XGBoost, SHAP | Demo | GitHub

- Built **end-to-end MLOps pipeline** with ZenML (9 steps: ingestion → validation → training → evaluation → registration)
- Achieved **88.52% PR-AUC** using XGBoost + LightGBM with Optuna tuning; SMOTE for imbalanced data
- Implemented MLflow tracking, DVC versioning, BentoML serving, SHAP + LIME explainability; Streamlit + CI/CD

QuantaAI – Agentic Search Assistant **2025**

LangGraph, GPT-4, Next.js 14, TypeScript, Tavily API | GitHub

- Built **4-node LangGraph state machine** (Classifier → Search → Generate → Response) with conditional routing
- GPT-4 routing achieving **94% accuracy**; Tavily web search; streaming with context management (10 messages)

DeepGuard – MLOps Deepfake Detection **2025**

TensorFlow, Xception, DVC, MLflow, AWS EKS, Grafana | Demo | GitHub

- Deepfake detection using Xception achieving **93.4% accuracy**; Grad-CAM explainability
- Production deployment on **AWS EKS Kubernetes** with Prometheus + Grafana monitoring; full CI/CD

VLMverse – Vision Language Models from Scratch **2025**

PyTorch, Transformers, SigLIP, Gemma, RoPE, KV-Cache | GitHub

- Implemented **PaLiGemma VLM**: SigLIP encoder (16×16 patches, 196 tokens) + Gemma 2B decoder-only LLM
- Built RoPE, KV-Cache for efficient inference, grouped-query attention, RMSNorm, GeLU activations

AgentForge – Multi-Agent Systems Framework **2025**

CrewAI, LangGraph, AG2 (AutoGen), smolagents | GitHub | 2 Stars

- **15+ agent implementations**: orchestrator, explorer, coder with tool use, function calling, memory systems
- Multi-agent collaboration patterns reducing complex task completion time by **60%**

Reasoning LLMs from Scratch **2025**

PyTorch, Llama 3.2, Chain-of-Thought, Beam Search | GitHub

- Implemented inference-time compute scaling: chain-of-thought prompting, beam search decoding strategies

Technical Skills

Machine Learning: Supervised Learning, Unsupervised Learning, Classification, Regression, Clustering, Feature Engineering
Deep Learning: PyTorch, TensorFlow, Keras, Neural Networks, CNNs, RNNs, Transformers, HuggingFace, Transfer Learning
Computer Vision: YOLO (v5-v8), OpenCV, Object Detection, Image Segmentation, ResNet, Xception, Vision Transformers
NLP/LLM: Text Classification, NER, Sentiment Analysis, BERT, GPT, Text Embeddings, Semantic Search
Generative AI: LangChain, LangGraph, LlamaIndex, RAG, Prompt Engineering, Fine-tuning, LoRA/QLoRA, PEFT, CrewAI
AI Agents: Multi-Agent Systems, Tool Use, Function Calling, Agentic Workflows, State Machines, Memory
MLOps: ZenML, MLflow, DVC, BentoML, Optuna, Model Registry, Experiment Tracking, A/B Testing, Docker, Kubernetes
Cloud/Tools: AWS (EC2, S3, EKS), GCP, Firebase, Netlify, FastAPI, Streamlit, CI/CD, Prometheus, Python, SQL, MongoDB, Pinecone

Education

B.Tech in Artificial Intelligence and Data Science **Nov 2022 – May 2026**
Lakshmi Narain College of Technology, Bhopal CGPA: 7.2/10
Coursework: Machine Learning, Deep Learning, Computer Vision, NLP, Reinforcement Learning, Neural Networks

Certifications & Achievements

ML Specialization - Udemy | **Generative AI** - Google Cloud | **Open Source:** 47 repos, 40+ stars, 10+ PRs merged | **CNCF Member**