

VENKATESH R

AI Engineer

hessiancurve@gmail.com | +91 8668135371 | LinkedIn | GitHub | Portfolio | Dharmapuri, India

AI Engineer with production experience building LLM-powered NLP pipelines, multimodal computer vision systems, and autonomous LangGraph agents deployed for Indian government agencies. Shipped 11 production NLP packages, built face/voice/text embedding pipelines processing forensic and surveillance data at scale, and cut manual report generation from days to minutes. 6x hackathon winner, ₹30L+ in prizes.

TECHNICAL SKILLS

LLM Frameworks	LangChain, LangGraph, DSPy (prompt optimization), LlamaIndex, Pydantic (structured outputs), RAG, MCP
Local LLMs & Inference	Gemma 3 12B, Llama 3, GLM, Phi-3, Ollama, vLLM, Hugging Face Transformers, quantization (GGUF/AWQ)
Deep Learning	PyTorch, TensorFlow, backpropagation, neural network design, transformer architectures, RLHF
Classical ML	Random Forest, XGBoost, LightGBM, SVM, Isolation Forest, DBSCAN, reinforcement learning (RL)
Vision & Audio	Florence-2, InsightFace, VGGFace2, YOLO (Ultralytics), PaddleOCR, Tesseract, Seamless-M4T, Whisper, NeMo
Embeddings & Retrieval	Qdrant, sentence-transformers (384-dim), BGE-M3 (1024-dim), cosine similarity search, MLflow
Languages	Python, Go, Rust, Bash

PROFESSIONAL EXPERIENCE

AI Engineer (Consultant) — Saptang Labs, Chennai (Jun 2025 – Jul 2026)

Altair — Forensic Intelligence Platform (*flagship — sole engineer, built end-to-end*) — designed and built a production AI-powered forensic intelligence platform integrating LLMs, computer vision, OCR, and semantic retrieval to accelerate large-scale document analysis.

- GPU inference worker** — Phi-3 NER for entity extraction, BGE-M3 1024-dim embeddings (Qdrant), and Neo4j graph writes for multi-hop link analysis; unified keyword/semantic/graph retrieval behind one API.
- Multilingual document intelligence** — 16+ document formats, structure-preserving extraction (PyMuPDF+OCR, calamine, LibreOffice) feeding downstream LLM-based analysis.

Coastal Surveillance — enhanced an enterprise multimodal AI platform by evaluating and integrating production inference pipelines for Florence-2, InsightFace, PaddleOCR, Tesseract, and Seamless-M4T, enabling large-scale image, video, document, and audio intelligence.

- Embedding-based identity search** — InsightFace → 512-dim VGGFace2 embeddings → DBSCAN clustering (cosine 0.65) → Qdrant, **cutting investigator search time from days to seconds** across massive multimedia datasets.
- Engineered scalable AI inference/orchestration for asynchronous multimodal processing using vLLM, Ollama, and distributed GPU task execution — improving reliability and throughput for production deployments.

Nucleus — built and deployed **11 reusable LLM-powered AI services** as installable Python wheels (typed Pydantic interfaces, dual Ollama/vLLM backends) using Gemma 3 12B, Phi-3, and DSPy — NER, relationship extraction, sentiment (80% accuracy), summarization, translation (14 languages), event extraction, narrative generation.

Report Generation Agent (LangGraph) — **cut analytical report generation from 3–5 days to ~25 minutes** by building agents that retrieve, dedupe (Qdrant similarity), rank, and synthesize 94+ articles/run into Jinja2-templated reports; added SQLite session memory and shipped as a cron-scheduled wheel.

Machine Learning Engineer — Google Student Developer Clubs, IIT Madras (Mar – Aug 2024, Part-time)

Built supervised ML models end-to-end (Linear/Logistic Regression, Decision Trees, Random Forest, KNN, SVM, XGBoost) — preprocessing, feature engineering, cross-validation, hyperparameter tuning, bias-variance analysis.

PROJECTS

Glofsense — LSTM + XGBoost GLOF prediction from SAR imagery + IoT sensors. **Winner, C-DAC HimaShield 2025, ₹5L**

Xylem — ZK-SNARK privacy-preserving KYC, zero PII exposure, 500ms proofs. **Winner, Cyber Security Innovation Challenge 1.0, ₹10L — Team Eigen**

Lamina — Autonomous AI agent for tokenized RWA lifecycle on Hedera. **3rd Place, Hedera Apex, \$8K**

Other wins: NFSU 2023 Winner (₹5L) | Nirnay — NFSU 2025 Winner (₹1L) | Eigencurve — Army Terrier Cyber Quest 3rd (₹25K) | Hedera DLT Winner (\$9K) | BRICS 2nd Global

EDUCATION

Degree	Institution	GPA	Duration
MBA Financial Management	IGNOU	—	2026 – 2028
B.Tech Biotechnology	Bannari Amman Institute of Technology	7.78	2022 – 2026

Cross-Domain: CS50 AI (Cambridge) · Banking & Finance (IIT Kanpur) · Supply Chain Management (IIT Kanpur)