

Nirmit K. Tripathii

AI / ML ENGINEER & RESEARCHER

Production-grade AI systems — multi-agent LLM orchestration, multimodal RAG, speech & video ML, legal AI, and civic tech. M.Tech in Modeling & Simulation with 3+ years shipping from ideation to deployment on real-world (often free-tier) hardware.

Email nirmitktripathii@yahoo.com Location [Delhi NCR, India](#) · Remote Portfolio nirmitktripathii.github.io/PortfolioSite Interactive Lab Live Simulators ⚡
GitHub github.com/nirmitktripathii LinkedIn [in/nirmitk-tripathii-553a92158](#)

AVAILABILITY

Open to freelance, consulting & research collaboration — remote. Open to full-time from Sept 2026.

CORE SKILLS

Agentic AI & LLMs

LangGraph LangChain LangSmith MCP
RAG · LightRAG · CRAG VideoRAG FAISS

Languages & Frameworks

Python TypeScript / Node PyTorch
FastAPI Flask SQL

Models & APIs

Gemini OpenAI Llama 3.1 HuggingFace
IndicF5 XLM-RoBERTa IndicBERT
InLegalBERT ImageBind Whisper

Data & MLOps

PostgreSQL Neo4j NetworkX HNSWLib
Docker W&B OpenCV Azure OCR
Pandas · NumPy

Sports AI

xG Modelling VAEP FUTR T-DEED
SoccerNet

EDUCATION

M.Tech — Modeling & Simulation

Savitribai Phule Pune University
Sep 2019 – Feb 2022

B.Tech — Electrical Engineering

Pandit Deendayal Petroleum University
Jul 2014 – Jun 2018

CERTIFICATIONS

Masters Program — Data Scientist

Simplilearn

Football Data Analytics

MAS About Sports

PUBLICATIONS & WRITING

11 engineering deep-dives, white papers & field notes across dubbing fine-tuning, multi-agent MCP, SoccerNet CV & civic systems. See [Page 2](#) for full catalog.

EXPERIENCE

AI/ML Software Engineer · Chetu India

Apr 2025 – Jun 2026 · Noida

- Built **DIGEST** — an agentic LangGraph (StateGraph) chatbot for event discovery & ticket commerce, orchestrating multi-step workflows across 7 intents with the full Automatiq B2B transaction lifecycle.
- Architected a **Video-to-SOP** multimodal pipeline: ImageBind embeddings + a Neo4j/NetworkX VideoRAG knowledge graph, a LangGraph refinement loop, and LangSmith observability.
- Built **IKS Oracle (Gyanstrot)** — a RAG system over Indian Knowledge Systems, deployed live on HuggingFace Spaces [\[Repo\]](#).

AI/ML Engineer · Blu Parrot

Jan 2024 – Apr 2025 · Gurgaon

- LegalGPT** — full RAG pipeline for Indian legal research & compliance; fine-tuned InLegalBERT/InCaseLawBERT across 150+ contract types; compared Naive/Light/CRAG retrieval.
- Email Orchestrator** — a multi-account MCP email agent consolidating 4 providers behind one MCP server that is itself an MCP client; 17 unified tools, per-account fan-out resilience, 49 passing tests [\[Build Log\]](#).
- Hindi ASR portal on NeMo Conformer-CTC (**92%+** zero-shot) with Gemini summarisation; a 22-language Indic translation portal (IndicBERT + IndicTrans2, **94%** zero-shot).
- Promoted to **Project Lead** for a UK real-estate multimodal appraisal regression model [\[Inferred Theorem\]](#).

Subject Matter Expert, AI/ML · House of Coutons

Nov 2023 – Jan 2024

- Delivered 3 research papers and a dissertation across literature review, EDA, model building and inference.

AI/ML Programmer · HRBOTICS (Credence Maven)

Apr 2022 – Mar 2023

- Multi-threaded Python scrapers for NCR commercial data; automated LinkedIn hiring workflows (Selenium + BeautifulSoup) with a reusable ATS prototype.

Football Data Analyst · Boomer 11

Nov 2021 – Mar 2022

- Built an **Expected Goals (xG) model** (RF Classifier) at **89.77%** accuracy — published research [\[Live Simulator\]](#); delivered player-rating and real-time performance analysis.

SELECTED PROJECTS

Indic AI Dubbing Platform · Llama-3.1-8B QLoRA · IndicF5 · Gemini · Faster-Whisper · Demucs [[Repo](#) · [Walkthrough](#) · [Build Log](#)]

Isochrony-aware dubbing across 11 Indic languages. The v3 fine-tune constrains translation to a phoneme budget, cutting length error **79%** (0.495→0.103 rel.) while *improving* quality (chrF++ 22.0→30.1); deployed v2 uses Gemini CoT + IndicF5 zero-shot cloning on free Kaggle T4.

FAANTRA — SoccerNet 2026 Ball Action Anticipation · PyTorch · FUTR · T-DEED · RegNetY-008 [[Research Note](#)]

Predicts ball actions seconds ahead from broadcast video with decoupled what/when heads. Official challenge-server **mAP_Average 17.69** (mAP@∞ 23.59, mAP@5 22.47) — decomposed metric identifies temporal boundary limits.

India Aaspaas — Civic Reporting Platform · Node/Express · PostgreSQL · React · Cloudflare R2

[[Live App](#) · [Repo](#) · [White Paper](#)]

Production map-based civic-issue platform with server-rendered social share cards, two-tier trust model, and hardened upload path — 67 tests against real Postgres engine, ~\$3/month to run.

Podcast Shorts Factory · Multi-agent · Python · open source (MIT) [[Repo](#) · [White Paper](#)]

Ten cooperating AI agents turning long podcasts into scheduled short-form videos — provider-agnostic LLM layer with automatic fallback, learning feedback loop, and QA gate; runs unattended on free providers.

AI Video Style Cloner · Streamlit · Multi-provider LLM failover · Edge-TTS · MoviePy [[White Paper](#)]

Channel URL → reverse-engineers house style, writes original script in that voice, narrates & compiles finished MP4 for **\$0**. 3-tier LLM failover, self-healing JSON validators, and zero-network procedural fallback.

Football Analytics — VAEP & Expected Goals · RF · SVR · Streamlit [[Live Space](#) · [Repo](#) · [Simulator](#)]

VAEP action-valuation dashboard across Top-5 leagues & World Cup 2018; published xG thesis (RF Classifier, 89.77% / 90.55% post-HPO).

11 full-length technical papers, architecture build logs, and empirical field notes from production deployments. Available online at nirmitktripathii.github.io/PortfolioSite/#writing

Portfolio nirmitktripathii.github.io/PortfolioSite **Interactive Lab** [Live Simulators](#) ⚡ **GitHub** github.com/nirmitktripathii **LinkedIn** in/nirmitk-tripathii-553a92158
Email nirmitktripathii@yahoo.com

Field Notes & Empirical MLOps (Straight from the Repo)

3 Articles

Teaching an LLM to Fit Its Mouth

12 min · 2026

[Fine-Tuning](#) [QLoRA](#) [Isochrony](#) [Llama 3.1](#)

AI dubbing drifts because translations differ in syllable count. Solved in the translator layer via fine-tuned Llama-3.1-8B QLoRA with phoneme budgeting — cut length error from 49% to 10% with full training telemetry.

[Read Deep-Dive](#) →

A Headless GPU Lab on Kaggle's Free Tier

9 min · 2026

[MLOps](#) [GPU Tunneling](#) [Automation](#)

Zero-cost remote GPU lab: secure reverse tunnel, mobile webhook trigger, public message feedback, and emergency kill switches. Details the PyTorch meta-tensor loading bug that caused silent failures.

[Read Deep-Dive](#) →

Predicting What Happens Next in Football Video: Action Anticipation

10 min · 2026

[Computer Vision](#) [SoccerNet 2026](#) [FUTR / T-DEED](#) [Action Anticipation](#)

Anticipating events seconds before they occur using video transformers. Analysis of official challenge-server mAP metrics (mAP@∞ 23.59 vs mAP@5 22.47) demonstrating decoupled what/when temporal classification failure modes.

[Read Deep-Dive](#) →

The Email Orchestrator · Multi-Account MCP Agent Build Log

3-Part Series

Part 1 · One Inbox, Three Mailboxes, Zero New APIs

12 min · 2026

[MCP Architecture](#) [Fan-out](#) [17 Unified Tools](#)

Consolidating Gmail, Zoho, Yahoo & Outlook behind a single MCP server that acts as an MCP client. Architecture, unified schema routing, and multi-tenant tool registry.

[Read Part 1](#) →

Part 2 · One Dead Inbox Shouldn't Kill the Query

11 min · 2026

[Resilience](#) [Process Isolation](#) [Promise.allSettled](#)

Handling child process crashes in MCP stdio connections: graceful partial degradation, self-healing reconnection pools, drop detection, and 49 passing unit/integration tests.

[Read Part 2](#) →

Part 3 · The Model Spent Its Whole Budget Thinking

11 min · 2026

[LLM Reasoning](#) [Thinking Models](#) [Token Budget Scaling](#)

Debugging silent JSON truncation in reasoning models: why small tasks fail with standard token limits when internal chain-of-thought consumes the full budget, and the dynamic token floor formula to fix it.

[Read Part 3](#) →

White Papers & Production Systems Architecture

5 White Papers

Ten Agents, One Newsroom: Video Factory

13 min · Aug 2026

[Multi-Agent](#) [LLM Orchestration](#) [QA Gate](#)

Ten cooperating agents turning long audio into scheduled short-form video: provider-agnostic fallback layer, craft playbooks, learning loop, and strict QA output gates.

[Read White Paper](#) →

Never Return Empty: Zero-Cost Video Pipeline

12 min · Aug 2026

[LLM Failover](#) [Self-Healing](#) [\\$0 Stack](#)

Zero-cost video generator: 3-tier multi-provider failover cascade, self-healing JSON schema repair, and zero-network procedural fallbacks ensuring guaranteed video delivery.

[Read White Paper](#) →

Single Origin: Production Civic-Reporting

12 min · Aug 2026

[Full-Stack SSR](#) [~\\$3/mo Budget](#) [Security](#)

Production civic-issue mapper: server-rendered dynamic social cards for WhatsApp crawlers, two-tier anonymous trust model, and hardened image sanitization against GPS leakage.

[Read White Paper](#) → [Live App](#) ↗

The Deployed Dubbing Pipeline (v2 System)

11 min · 2026

[Speech ML](#) [Demucs DSP](#) [IndicF5 Cloning](#)

Production v2 dubbing system: phoneme-budgeted translation with Gemini CoT, Demucs 4-stem soundstage preservation, and duration-controlled voice cloning on a free Kaggle T4.

[Read White Paper](#) →

The 230-Year Constraint: India's Mechanization & Draft Power Problem

16 min · Aug 2026

[Agri-Policy](#) [Economic Modeling](#) [Primary Historical Sources](#)

Historical data analysis & policy research analyzing India's agricultural mechanization bottleneck from 18th-century primary sources to modern fractional horsepower constraints.

[Read White Paper](#) →