

HARSH TOMAR

AI / GenAI Engineer

+91-8817969476 • tomarharsh28303@gmail.com • [LinkedIn](#) • [GitHub](#) • [Portfolio](#)

SUMMARY

AI/ML engineer with **2+ years of hands-on system building** (2024–2026) across production LLMOps, multi-agent architectures, and LLM evaluation pipelines. Built **Continuum**, an 11-microservice LLMOps platform with Kafka trace capture, custom LLM-as-judge scoring, and SPC-based drift detection on AWS ECS Fargate. Built and deployed a RAG service at **i3 Digital Health**; built **QuantaAI**, a LangGraph multi-agent system with a hand-labelled eval set (94% routing accuracy); fine-tuned **Qwen 2.5 1.5B** via QLoRA; reimplemented **PaLiGemma VLM** in raw PyTorch. Hands-on with LangChain, LlamaIndex, vector DBs (Pinecone, Chroma, FAISS, Weaviate), and OpenAI / Anthropic / Gemini APIs.

EXPERIENCE

AI Intern

May 2025 – Jul 2025

i3 Digital Health

Remote

- Built and deployed a **RAG pipeline for researcher profiling**, aggregating and cleaning data from PubMed, Google Scholar, and ResearchGate via multi-source APIs and scraping; **LangChain** retrieval over vector embeddings drove contextual collaborator recommendations
- Owned the full build-to-deploy loop: iterated chunking strategies, embedding models, and retrieval prompts with healthcare SMEs; when their judgment on trust diverged from my retrieval metrics, rebuilt the system around source attribution and a refuse-when-unsure rule

PROJECTS

Continuum: LLMOps Platform for Quality Monitoring & Self-Healing

2026

AWS ECS Fargate, Apache Kafka, FastAPI, React/TypeScript, Terraform, PEFT, MLflow, Prometheus | [GitHub](#)

- **Problem:** detect LLM quality degradation in production traffic and orchestrate fine-tuning to remediate it
- **Approach:** **11-microservice event-driven platform** on AWS ECS Fargate with an **Apache Kafka** trace pipeline; multi-dimensional quality scoring (Claude LLM-judge + ROUGE/BLEU + semantic similarity); **Statistical Process Control** (Western Electric rules) for drift detection; designed a self-healing loop for **LoRA** fine-tuning with A/B canary deployment; full **Terraform IaC** and GitHub Actions CI/CD
- **Result:** verified end-to-end trace flow (Kafka to scoring to PostgreSQL) with **166 Prometheus metric series**, an authenticated REST API, a React dashboard, and 10/10 passing E2E tests

QuantaAI: LangGraph Search-Routing Chatbot

2025

LangGraph, GPT-4, Tavily Search, FastAPI, Next.js 14, TypeScript | [GitHub](#)

- **Problem:** chat assistant that decides on its own when a query needs live web search vs. direct generation
- **Approach:** **4-node LangGraph state machine** (Classifier → Search → Generate → Response) with a GPT-4 routing classifier; **Tavily** web-search integration with relevance ranking; SSE streaming via FastAPI to a Next.js 14 frontend; sliding 10-message context window
- **Result:** **94% routing-classifier accuracy** on a hand-labelled eval set; transparent multi-stage UX exposing classification, search, and generation phases

VLMverse: PaLiGemma Vision-Language Model

2025

PyTorch, SigLIP, Gemma 2B, RoPE, KV-Cache, Grouped-Query Attention, RMSNorm | [GitHub](#)

- **Problem:** internalise modern VLM architecture by reimplementing one end to end
- **Approach:** reimplemented **PaLiGemma** in pure PyTorch following the paper and reference walkthroughs: SigLIP encoder (16×16 patches to 196 tokens), **Gemma-2B** decoder, **RoPE** positional embeddings, **KV-Cache** for autoregressive inference, grouped-query attention, RMSNorm
- **Result:** working end-to-end inference; deep working knowledge of attention, positional encodings, and cross-modal projection

ClinicalQwen: Medical Q&A Fine-Tuning on Qwen 2.5 1.5B

2026

PyTorch, Qwen 2.5 1.5B, QLoRA, 4-bit NF4, PEFT, Hugging Face | [GitHub](#)

- **Problem:** adapt a general-purpose LLM to answer clinical and medical queries accurately
- **Approach:** QLoRA fine-tuning of Qwen 2.5 1.5B on 5,000 medical Q&A samples (lavita/medical-qa-datasets); 4-bit NF4 quantisation; PEFT with LoRA rank-8 adapters; trained on Google Colab T4
- **Result:** domain-adapted model with ~1% trainable parameters (9.2M / 1,552M); 86% keyword coverage, perplexity 3.14; deployed on Hugging Face Hub

Parallax: Self-Organising Knowledge Graph & MCP Server

2026

D3.js, MCP, Python, Full-Text Search | [GitHub](#)

- **Problem:** a growing personal knowledge base of projects, papers, and concepts becomes unsearchable and disconnected without structure
- **Approach:** 38 interlinked pages connected by a D3.js force-directed graph (33 nodes, 155 edges), colour-coded by domain; an MCP server exposes 5 callable tools (search_wiki, get_page, get_related, and others) so any MCP-compatible agent can query the graph directly
- **Result:** a working knowledge system that is both human-browsable and agent-queryable through the same underlying structure

Additional Work

- **issue-resolver-agent (2026):** autonomous GitHub issue fixer; SOUL.md behavioural spec + RULES.md hard constraints; escalation rules stop the agent when ambiguous or high-risk [\[GitHub\]](#)
- **Transformers-CV (173 commits, 2024–2026):** PyTorch reimplementations of ViT, DETR, Swin, SAM, Flamingo VLM, DDPM, JEPA, VAE/VQ-VAE from scratch [\[GitHub\]](#)
- **RAGCraft (2024):** a hub for RAG patterns, including advanced chunking, hybrid search, and re-ranking across Weaviate, Pinecone, FAISS, and Chroma [\[GitHub\]](#)
- **100-Days of Inference Engineering:** KV-Cache, speculative decoding, quantisation (FP8/INT8/INT4), vLLM, TensorRT-LLM [\[GitHub\]](#)

WRITING

- **Statistical Process Control for LLM Quality Monitoring:** Continuum's drift-detection engine, covering WECO rules, EWMA, and three honest gaps in the implementation [\[Read\]](#)
- **Deconstructing Autoresearch:** a technical teardown of Andrej Karpathy's autoresearch repo, covering attention patterns, scaling behaviour, and optimiser internals [\[Read\]](#)

OPEN SOURCE CONTRIBUTIONS

3 merged PRs to upstream projects:

- [BBoxMaskPose #12](#): Docker support for the ICCV 2025 paper repo
- [multimodal-agents-course #23](#): fixed a broken Kubrick UI build
- [hive #6849](#): refactored a deprecated storage API

TECHNICAL SKILLS

LLM & GenAI: LangChain, LangGraph, LlamaIndex, RAG, prompt engineering, agent orchestration, OpenAI / Anthropic / Gemini APIs, Hugging Face Transformers

Fine-Tuning & Architecture: LoRA, QLoRA, PEFT, 4-bit NF4 quantisation, attention mechanisms, RoPE, KV-Cache, grouped-query attention, RMSNorm

Vector & Retrieval: Pinecone, Chroma, FAISS, embedding models, semantic search, chunking strategies, re-ranking

Deployment & LLMOps: FastAPI, Docker, Kubernetes, AWS (ECS Fargate, EKS, EC2, S3, ECR), Apache Kafka, Terraform, MLflow, Prometheus, Grafana, GitHub Actions CI/CD, Hugging Face Spaces

Foundations: PyTorch (primary), TensorFlow, NumPy, Pandas, transformers, autoregressive decoding |

Languages: Python, SQL, TypeScript, Bash

EDUCATION

B.Tech in Artificial Intelligence & Data Science

Nov 2022 – May 2026

Lakshmi Narain College of Technology, Bhopal

CGPA: 7.37/10

Coursework: NLP, DL, ML, Reinforcement Learning, Linear Algebra, Probability, Statistics, Neural Networks