HARSH TOMAR

AI / GenAI Engineer

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

SUMMARY

AI engineer building **LLM-powered systems**: RAG pipelines, LangGraph agents, and parameter-efficient fine-tuning. Shipped a production RAG service for healthcare research at **i3 Digital Health**; built **QuantaAI**, a LangGraph state-machine chatbot with web-search routing; and reimplemented **PaLiGemma VLM** (SigLIP + Gemma-2B with KV-Cache and RoPE) in PyTorch for architectural depth. Hands-on with LangChain, LlamaIndex, vector DBs (Pinecone, Chroma, FAISS), and OpenAI / Anthropic / Gemini APIs.

EXPERIENCE

AI Intern **May 2025 – Jul 2025**

i3 Digital Health *Remote*

- Built a **RAG-powered researcher-profiling system** aggregating PubMed, Google Scholar, ResearchGate via multi-source APIs; **LangChain** retrieval over vector embeddings for contextual collaborator recommendations
- Iterated retrieval prompts and chunking strategies with healthcare SMEs; deployed via **FastAPI + Docker**

PROJECTS

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-labeled 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 + reference walkthroughs — SigLIP encoder (16×16 patches → 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

PyTorch LoRA & QLoRA – Parameter-Efficient Fine-Tuning **2025**

PyTorch, Low-Rank Adaptation, 4-bit NF4 Quantisation, PEFT | GitHub

- **Problem:** fine-tune large models on consumer GPUs without full-parameter updates
- **Approach:** pure PyTorch **LoRA** ($W = W_0 + BA$, rank 8/16/32) injected into attention layers; **QLoRA** with **4-bit NF4** + double quantisation; benchmark across BERT and LLaMA scales
- **Result:** $<$1% trainable parameters with maintained quality; $\sim$**65% memory reduction** (BERT) and up to 85% (LLaMA-65B)

Multi-Document RAG – PDF Chat over Embeddings **2025**

LangChain, OpenAI, Streamlit, FAISS, Vector Embeddings | GitHub

- **Problem:** conversational Q&A over a user-uploaded PDF corpus with source attribution
- **Approach:** chunking + embedding + FAISS retrieval; LangChain prompt orchestration with OpenAI; Streamlit chat UI with source citations
- **Result:** multi-doc semantic search with context-aware responses and inline source links

Additional GenAI Work

Transformers-CV (**173 commits**) — PyTorch reimplementations of **Flamingo VLM, DDPM diffusion, ViT, JEPA, VAE/VQ-VAE, GANs** [GitHub] • **RAGify** — collection of RAG patterns and pipelines • **Tennis Vision** (**31 stars**) — 30 FPS multi-model production CV pipeline

Open Source Contributions

3 merged PRs to upstream projects: BBoxMaskPose #12 (Docker support, ICCV 2025 paper repo) • multimodal-agents-course #23 (Kubrick UI build fix) • hive #6849 (deprecated storage refactor)

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: FastAPI, Streamlit, Docker, Hugging Face Spaces, AWS (EC2, S3, EKS), CI/CD
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: 7.2/10
Coursework: NLP, Deep Learning, Reinforcement Learning, Linear Algebra, Probability & Statistics, Neural Networks