

Hetansh Waghela

[+91-8369107794](tel:+91-8369107794) | hetanshwaghela06@gmail.com | [LinkedIn](#) | [GitHub](#)

EDUCATION

B.Tech in Computer Science and Engineering (Data Science) — CGPA 9.34 / 10
Dwarkadas J Sanghvi College of Engineering

Mumbai
2024 – 2028

EXPERIENCE

AhumLabs
AI Engineer Intern

Mumbai (Hybrid)
Mar 2026 – Jun 2026

- Built RAG for a voice/chat assistant on a multi-tenant trading platform, pulling tenant knowledge and past transaction history so users could fill bunker forms and look up vessel/pricing details in natural language.
- Shipped admin tooling for tenant onboarding, form customization, and knowledge uploads; each client's form fields and retrieval context feed directly into the assistant's tool schemas and prompts.
- Owned the embedding and indexing stack (local ONNX + Atlas Vector Search), with auto-indexing on new transactions and backfill scripts to bring existing data online without blocking live writes.

TECHNICAL SKILLS

Machine Learning: Supervised & Unsupervised Learning, Representation Learning, Model Evaluation, Error Analysis
Deep Learning: Transformer Architectures, Attention Mechanisms, Fine-Tuning (LoRA), Probing & Interpretability
NLP: Question Answering, Retrieval-Augmented Generation (RAG), Text Classification, Embeddings
Frameworks: PyTorch, Hugging Face Transformers, LangGraph, FastAPI, Pydantic, Instructor
Data & Tooling: NumPy, Pandas, Neo4j, ChromaDB, Mem0, Git, Docker

PROJECTS

ANVIL – Compliance-Grade RAG Over ASME Standards 🧠 – *Python, Pydantic, NetworkX, sqlite-vec, FastAPI*

- Built a compliance-grade RAG system for pressure-vessel queries with auditable provenance, achieving 0.950 pass rate and 1.000 calculation correctness on a 100-example SPES-1 benchmark with NVIDIA NIM.
- Engineered a hybrid retrieval stack combining BM25, sqlite-vec embeddings, and a typed NetworkX knowledge graph via RRF and reranking, reaching 1.000 retrieval recall@10.
- Shipped a bounded agentic tool loop with transcript logging and guardrails; diagnosed loop failure modes and improved agent pass rate from 0.460 to 0.640.

MYTHFORGE – Agentic Mythology Engine 🧠 – *Python, Gemini, LlamaIndex, Gradio, Neo4j, Mem0*

- Architected a multi-phase generative pipeline transforming a single seed concept into a fully interconnected mythology using Gemini, LlamaIndex RAG, and Pydantic-validated entities.
- Built an MCP server (7 tools) enabling external AI agents to create, query, and expand mythologies programmatically.
- Engineered graph-backed memory and retrieval with Mem0 and Neo4j to maintain cross-entity consistency and long-horizon coherence across generated worlds.

ImmunoPath – Immunotherapy Decision Support Pipeline 🧠 – *Python, PyTorch, LoRA, Hugging Face, Gradio*

- Built a 4-model clinical decision-support pipeline combining MedGemma, TxGemma, Path Foundation, and MedSigLIP to predict 8 immune biomarkers directly from routine H&E slides.
- Curated 60,396 patches across 950 TCGA NSCLC patients and diversity-selected 7,574 via K-Means for LoRA fine-tuning of a 4B vision-language model on A100-SXM4-80GB GPUs.
- Improved structured prediction reliability to 100% JSON/schema compliance and reduced TIL MAE to 0.16 on 94 held-out patients through response-only loss masking.

Ghosts in the Model – Machine Unlearning & Interpretability 🧠 – *Python, PyTorch, Transformers, GPT-2, LoRA*

- Evaluated Gradient Ascent and NegLoRA unlearning on GPT-2 over 59 landmark→city forget prompts, reducing forget probability to 7.6% and 0.8%, respectively.
- Designed Lazarus activation patching to restore 89.7–100% of forgotten behavior, showing knowledge was suppressed rather than erased despite decodable residual representations in linear probes.

ACHIEVEMENTS

- 4th place (Creative, Track 2) at Anthropic & Gradio's **MCP 1st Birthday Hackathon** with 7,000+ global participants.
- Second Runner-Up at **TechnoGrad 3.0** (DJSCE, Mumbai).
- Top 40 rank in **DataHack 3.0**, Mumbai's largest datathon.
- Open-source contributor to Stanford CRFM's **HELM** and **TransformerLens**.

EXTRACURRICULARS

DJS S4DS
Co-Committee Member (Machine Learning)

Mumbai
Sept 2025 – May 2026

- Designed ML/NLP problem statements for hackathons and workshops; built dataset curation and automated evaluation pipelines.