

Udit Rawat

7303993773 New Delhi, India
uditrawat1905@gmail.com GitHub LinkedIn

SUMMARY

AI Engineer with production experience building RAG pipelines, fine-tuning and deploying LLMs (LoRA/QLoRA) on edge hardware, and shipping multi-agent systems at scale. Springer-published on LLM adaptation; promoted from intern to engineer in under 2 months.

TECHNICAL SKILLS

Languages	Python (async), TypeScript, SQL, Shell
AI & ML	PyTorch, HF Transformers, LangChain, LangGraph, DSPy, RAG (pgvector, Qdrant), LLM Fine-tuning (LoRA/QLoRA), RAGAS
Model Optimization	GGUF, llama.cpp, ONNX, INT8/NF4 Quantization, Knowledge Distillation
Cloud & Backend	GCP, Azure, AWS, Cloudflare Workers/Pages, FastAPI, Docker, Kubernetes
MLOps & Data	LiteLLM, Langfuse, ClickHouse, PostgreSQL, CI/CD

EXPERIENCE

AI Engineer, TapHealth	New Delhi, India	Apr 2025 – Present
------------------------	------------------	--------------------

- Built and shipped an offline diabetes-guidance SLM running fully on-device on Snapdragon 680 hardware, in partnership with Eli Lilly: 3× smaller at 380 MB, under 890 ms latency at 25 tok/s, 98% accuracy on out-of-distribution refusal. Fine-tuned Qwen 0.6B/1.5B with LoRA, quantized to GGUF Q4_K_M, and built a separate ONNX INT8 MPNet classifier for OOD detection.
- Built a stateful 5-agent coaching system with LangGraph and Mem0/Hindsight memory across 25 tools, served via FastAPI with hybrid retrieval over pgvector/Qdrant. Built an async OCR/PDF pipeline streamed over SSE WebSockets, cutting report processing from 40s to 15s.
- Cut LLM costs 54% by designing a centralized LiteLLM gateway routing requests across OpenAI, Anthropic, and open-source models for 65+ agentic workflows spanning 5+ microservices, with structured outputs via Instructor and full request tracing via Langfuse.
- Rebuilt a monolithic 812-line rules-based nudge pipeline into a decoupled DSPy-based GenAI architecture (habit classifier separated from content generator), plus a standalone summary-generation service, serving 5K+ paid and free users. Built a week-0/week-N onboarding planner and daily scheduler for weekly planning, and a ClickHouse-backed analytics endpoint aggregating glucose, HbA1c, and meal data, with full CI/CD.

AI Intern, TapHealth	New Delhi, India	Feb 2025 – Mar 2025
----------------------	------------------	---------------------

- Built ARIMA/SARIMA/LSTM forecasting models for patient engagement and generated synthetic clinical data using NHANES + Q-learning, with LIME-based explainability. *Promoted to AI Engineer after 2 months.*

Project Intern, Perception & Intelligence Lab, IIT Kanpur	Jul 2024 – Oct 2024
---	---------------------

- Curated a 200K+ image dataset and optimized diffusion-based inpainting models, achieving **11.72 dB PSNR** and **0.41 SSIM**.

PROJECTS

Hierarchical Reasoner (*Python, PyTorch*) | github.com/udit-rawat/hierarchical-reasoner
From-scratch PyTorch reproduction of the Hierarchical Reasoning Model (arXiv:2506.21734) with dual-module GRU and ACT Q-learning adaptive halting. Reached 100% maze-path accuracy by epoch 20 with 66.7% compute reduction versus baseline; validated across 8 controlled experiments with 49 pytest cases.

Tribunal (*Python, LangGraph, MCP*) | github.com/udit-rawat/tribunal
Adversarial multi-agent fact-checker: parallel Prosecutor/Defender agents argue over just-in-time RAG evidence (hybrid Wikipedia/web search, ONNX MiniLM embeddings, LanceDB), adjudicated by a Judge. A deterministic citation verifier rejects any quote not found verbatim in the source, eliminating fabricated citations without a second LLM call. **83% verdict accuracy** on a 12-claim golden set. Shipped as CLI, MCP server, and a streaming FastAPI UI with per-node telemetry; **3.2× faster** retrieval via concurrent search plus a verdict cache.

PUBLICATIONS

Efficient Adaptation of Lightweight LLMs with T-Free Tokenization Boosting Algorithms for Class Imbalance	ICAAI 2025, Springer SSWC 2025
--	-----------------------------------

EDUCATION

B.Tech in AI & ML, VIPS-TC, GGSIPU, New Delhi	2021 – 2025	GPA: 9.1/10
---	-------------	-------------