RUDRA PRATAP SINGH

AI/ML Engineer | LLMs, PyTorch Fine-Tuning, AI Agents and RAG

Bengaluru, India | rudra19867@gmail.com | linkedin.com/in/rudra-pratap-singh-5a0ab532a | github.com/rudraksh01-tech |
Portfolio: rudraksh01-tech.github.io

SUMMARY

Computer Science student (B.E., CGPA 8.0) who fine-tunes, evaluates and ships machine learning and LLM systems. Four live, tested AI projects, including a PyTorch model that handles **57%** of support tickets without an LLM at near-equal accuracy. Seeking AI Engineer and Machine Learning Engineer internships.

EDUCATION

Bachelor of Engineering (B.E.), Computer Science and Engineering | 2024 - 2028 (Expected)
Cambridge Institute of Technology, North Campus, Bengaluru (VTU) | CGPA: 8.0 / 10 (5th semester)
Coursework: Data Structures and Algorithms (DSA), OOP, Operating Systems, DBMS, Computer Networks

PROJECTS

Triage | Fine-Tuned Model + LLM Cascade for Support Tickets | Nov 2026 | Live demo: triage-rho-eight.vercel.app
Python, PyTorch, Hugging Face, scikit-learn, ONNX, FastAPI, BullMQ, Redis, Next.js | github.com/rudraksh01-tech/triage

- Cut LLM calls by **57%** at near-equal accuracy (**80.7%** vs 81.0% LLM-only; 384 held-out tickets labelled by a stronger LLM): a MiniLM fine-tuned in PyTorch on 18,000 LLM-labelled tweets answers in **8 ms** and escalates only unsure tickets.
- Beat a TF-IDF + logistic regression baseline (71% vs 63%); matched full fine-tuning with LoRA (0.7% of weights); cut calibration error 0.108 to 0.066 (temperature scaling); quantized to ONNX int8 (**4x faster**) after a parity check caught an export bug.
- Served it via FastAPI (**100 req/s** at 42 ms p95, 0 errors) fed by a BullMQ/Redis queue, streaming to a Next.js inbox over WebSockets; a circuit breaker kept tickets labelled when the LLM quota ran out; 16 pytest and 8 Playwright tests in CI.

Dwarpal | AI Security Gateway for LLM Agents (MCP) | Nov 2026 | Live demo: rudraksh01-tech.github.io/dwarpal
Python, LangGraph, MCP, FastAPI, PostgreSQL, pgvector, Redis, Docker, Kubernetes, Helm | github.com/rudraksh01-tech/dwarpal

- Cut data leaks from **5 of 25** attack runs to **0 of 25**, with no false blocks on attack-free runs, via an open-source gateway between LLM agents and MCP tool servers that stops prompt injection, tool poisoning and rug pulls in four languages.
- Architected a LangGraph policy graph with human-in-the-loop approval: SHA-256 tool pinning, cross-server taint tracking, multilingual rules, Llama Prompt Guard 2 and an LLM policy judge, cheapest checks first.
- Found that Prompt Guard 2 scored the one attack that fooled the agent **0.001** (safe) while the policy judge caught it; shipped Docker images and a Helm chart that CI tests on Kubernetes; **34** tests.

AyurIP-Sahayak | Multilingual RAG Assistant | Aug 2026 - Nov 2026 | Live demo: ayur-ip-sahayak.vercel.app
Python, FastAPI, React, Google Gemini API, NumPy, Docker, GitHub Actions | github.com/rudraksh01-tech/AyurIP-Sahayak

- Reached Hit@1 **0.95** and MRR **0.97** on 38 labelled questions in four languages with dense retrieval (Gemini embeddings, NumPy search) vs 0.47 and 0.60 for BM25; kept dense as the default after a BM25 + RRF hybrid scored lower (MRR 0.93).
- Validated answers with an LLM judge from another model family: **0.99** faithfulness (166 of 167 claims), 0.98 citation precision, and the judge caught **15 of 15** planted errors; first words stream in under 2 s; 47 tests in CI.

NestFinder | Agentic AI Assistant for Flat Search | Oct 2026 - Nov 2026 | Live demo: nestfinder-woad.vercel.app
Node.js, Express.js, React, MongoDB, Groq API, Langfuse, OpenTelemetry | github.com/rudraksh01-tech/nestfinder

- Raised tool-choice accuracy from **18/20 to 20/20** with a 20-case eval (Hindi, Hinglish and prompt-injection inputs) of a tool-calling agent built from scratch; traced every LLM call with OpenTelemetry and Langfuse (PII masked); 26 unit tests in CI.

EXPERIENCE

Founding Software Engineer Intern, FERMOR | Bengaluru (Hybrid) | Sep 2026 - Oct 2026

- Drove **about 30x** the usual daily traffic on launch day (a month's worth in one day) by redesigning key website pages end to end and optimizing them for search engines (SEO); measured the impact with web analytics.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, SQL, Java | **Frontend:** React (React.js), Next.js, Tailwind CSS

Machine Learning and Deep Learning: PyTorch, Hugging Face Transformers, Fine-Tuning, LoRA (PEFT), Knowledge Distillation, Text Classification, Model Evaluation, Error Analysis, Calibration, scikit-learn, Pandas, NumPy, ONNX, Quantization

Generative AI (GenAI) and LLMs: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), AI Agents, LangGraph (LangChain), Model Context Protocol (MCP), Tool Calling, Embeddings, Vector Databases (pgvector), Hybrid Search, LLM-as-a-Judge, Guardrails, Prompt Engineering, Natural Language Processing (NLP)

Backend, Deployment and Testing: FastAPI, REST APIs, WebSockets, Redis, BullMQ, PostgreSQL, MongoDB, Docker, Kubernetes, Helm, CI/CD (GitHub Actions), Git, Linux, pytest, Playwright, Langfuse