

Gaurav Kumar Nayak

Machine Learning Engineer | AI Systems Developer

Bhubaneswar, India · +91-8658995681 · gauravnayak711@gmail.com · LinkedIn · GitHub · Portfolio

TECHNICAL SKILLS

Languages	Python, SQL, TypeScript / JavaScript (ES2022), C++, HTML
CS Fundamentals	Data Structures & Algorithms (DSA), Object-Oriented Programming (OOP), System Design, Relational Databases, REST API Design, Git / Version Control, Linux
ML / Deep Learning	PyTorch, Natural Language Processing (NLP), LightGBM, XGBoost, Scikit-learn, SHAP, Feature Engineering, Graph Neural Networks (GNN: GraphSAGE, GATConv), Survival Analysis (Cox PH), Post-Training Quantization (AIMET), Qualcomm AI Hub (QNN/HTP), GPU Programming (OpenAI Triton, CUDA)
LLM & Agentic AI	Large Language Models (LLMs), LangGraph, LangChain, Retrieval-Augmented Generation (RAG), Vector Databases (ChromaDB, pgvector), HuggingFace Transformers, Fine-Tuning (QLoRA/Unsloth), Mechanistic Interpretability (TransformerLens), Agentic Tool Use
MLOps & Infra	FastAPI, Docker, Redis, PostgreSQL, Kafka, Model Monitoring (Evidently AI, Drift Detection), Prometheus, CI/CD (GitHub Actions), AWS (S3, Bedrock), Unit Testing (pytest), Microservices

EDUCATION

C.V. Raman Global University, Bhubaneswar

B.Tech, Computer Science (Data Science)

Sep 2024 – May 2028 (Expected)

CGPA: 7.8 / 10.0 (4th Semester)

TECHNICAL PROJECTS

CodeSageZ – Graph-Augmented RAG for Repository-Level Code Q&A [GitHub] Jun 2026 – Present

FastAPI, Next.js 14, PostgreSQL, ChromaDB, NetworkX, Tree-sitter, QLoRA/Unsloth, Docker, pytest

- Boosted code retrieval from **0% to 53.3% Recall@8** (95% CI: 44.4–62.0%, $n=120$ edges, p95 5.8 ms) by building a Retrieval-Augmented Generation (RAG) engine that expands vector seeds via 1-hop BFS over a Tree-sitter AST call graph and re-ranks with a hybrid score ($0.6 \times \text{cosine} + 0.4 \times \text{graph proximity}$).
- Improved NLP code generation by **+9.38 CodeBLEU pts (+15.5%, 60.64 → 70.02)** on 3.8 GB VRAM by fine-tuning Qwen2.5-Coder-1.5B via 4-bit QLoRA (Unsloth) on 8K CommitPack bug-fix commits with CodeBLEU + HumanEval evaluation harness.

HyperFlow – Hyperlocal Commerce Intelligence Platform [GitHub] Apr 2026 – Present

Tobit MLE, Cox PH, LightGBM Quantile, Redis Lua, PSI Drift Detection, FastAPI

- Recovered censored demand signal achieving **wMAPE 38.99% → 29.53% (–24.3%)** across 42K SKU time-series by formulating a heteroscedastic Tobit Maximum Likelihood Estimation (MLE) forecaster (Inverse Mills Ratio imputation) chained into LightGBM Quantile; PSI drift monitor auto-triggers retraining at $\text{PSI} > 0.20$.
- Reduced rider-trips by **62%** (50 orders → 19 batches in **10.4 ms**, 15-min SLA) by implementing a Haversine nearest-neighbour dispatch batcher backed by a Redis Lua atomic lock to eliminate dual-dispatch race conditions.

QualEdge – Qualcomm Snapdragon Edge AI Optimisation Suite [GitHub] Mar 2026 – Jun 2026

AIMET 2.34.0, Post-Training Quantization (PTQ), Qualcomm AI Hub (QNN/HTP), MobileNetV2, ModernBERT, FastAPI

- Compressed MobileNetV2 by **4.04×** (14.16 MB → 3.51 MB, <0.5% accuracy drop) and accelerated NPU inference by **41×** (22.78 ms → 0.55 ms on Snapdragon X Elite) by executing 4-stage Post-Training Quantization (PTQ) with 0 CPU fallback ops and 100% Hexagon HTP native execution.
- Achieved **93.3% hybrid LLM routing accuracy at 0.52 ms p50** across 600 benchmarks by training a TF-IDF + ModernBERT + Logistic Regression classifier to partition on-device vs. cloud inference dynamically; 20 pytest tests.

Aarthaa AI – FinTech Fraud Detection & Compliance [GitHub] May 2026 – Present

LightGBM, SHAP, ChromaDB, pgvector, LLaMA-3.1-70B, Evidently AI, Prometheus, FastAPI

- Detected payment fraud with **AUC-ROC = 0.914 and Recall = 61.7%** on a 118K-transaction holdout by training LightGBM on 590K IEEE-CIS records (temporal split, $\text{scale_pos_weight} = 27.5$) with SHAP global feature attribution on top-10 fraud signals.
- Achieved **93.3% top-3 compliance retrieval accuracy** across 20 RBI circulars by building a RAG agent (ChromaDB + pgvector + LLaMA-3.1-70B) with Evidently AI drift monitoring, Prometheus metrics, and a LangChain SQL agent for hallucination-free ledger queries.

RailMind – Autonomous Railway Dispatch Engine [GitHub] Jun 2026 – Present

LangGraph, Graph Neural Network (GNN: GraphSAGE/GATConv), XGBoost, Redis Streams, TimescaleDB, PostgreSQL

- Classified train delay cascades with **F1 = 0.908, AUC-ROC = 0.981, Precision = 0.933** on holdout ($n=3,000$) by training a Platt-scaled stacking ensemble (GB+RF+LR) on Graph Neural Network (GNN) embeddings inside a compiled LangGraph StateGraph.
- Guaranteed **100% tamper-proof** dispatch auditability at sub-150 ms SLA (**76% test coverage, 181 tests**) by designing a SHA-256 cryptographic audit ledger and a 4-tier data pipeline (NTES → PostgreSQL) with exponential backoff.

TritonForge – GPU Kernel Optimisation Workstation [GitHub] May 2026 – Present

OpenAI Triton, PyTorch, CUDA, GPU Programming, NVIDIA T4

- Achieved **4.47×** throughput speedup (1.70 ms → 0.38 ms) and **93.1% HBM utilization** (297.8 GB/s of 320 GB/s T4 peak) by fusing RM-SNorm + Linear QKV projection into a single OpenAI Triton kernel with SRAM transaction consolidation and $O(1)$ activation recomputation.
- Cut peak VRAM by **95.3%** (134.2 MB → 6.3 MB at seq 2048) and accelerated attention **2.65×** over PyTorch eager by implementing Tiled FlashAttention-2 with online softmax scaling in SRAM.

HACKATHONS & COMPETITIVE ACHIEVEMENTS

- Google Big Code Round 1** – Qualified; top 15,000 globally. (2026)
- Meta × OpenEnv AI Agent Challenge** – Top 800 / 31,000+ teams; PyTorch Reinforcement Learning (RL) env for automated support-ticket triage. (Apr 2026)
- Amazon Nova AI Hackathon** – Prize Winner; 5 Nova models on AWS Bedrock + FAISS for contradiction detection. (Jan 2026)
- Zaggle × COMET'26 IIT Roorkee** – Winner · **IIT Bombay Kaizen** – Winner · **ArtPark CodeForge IISc** – Winner. (2025–2026)

CERTIFICATIONS

Machine Learning (ML) Specialization – Stanford×DeepLearning.AI (Apr 2026) | Generative AI Engineering with LLMs – IBM×Coursera (Apr 2026) | AWS Cloud Practitioner (May 2026) | Meta Full Stack Developer (Apr 2026)