Jyotirmoy Konwar

Github | LinkedIn | jyotirmoykonwarjk@gmail.com | Portfolio | +91-6901262645

Experience

Siemens May 2026 - July 2026
Technical Intern *Bengaluru, KA*

- Architected an **autonomous agentic code-repair Python package** leveraging AST traversal and dependency graphs, synthesizing verifiable production patches to resolve CI/CD build failures.
- Integrated a **multi-stage validation harness** (unit tests, property-based testing, formal checks) to verify patch soundness before automated staging, testing, and production rollout.
- **Architected a LLM benchmarking and evaluation framework** using **rubric-based scoring** for qualitative and quantitative evaluation of **automated legacy code migration** agents.

Freelance Project August 2025 - September 2025
Freelance Project for a Client *Remote*

- Engineered a **low-latency**, **zero-shot multimodal** document parsing pipeline integrating **Qwen-2.5-VL with 4-bit GGUF quantization (llama.cpp)** for high-efficiency CPU inference.
- Achieved **98% extraction accuracy** and an **8× reduction in memory footprint** across **high-throughput batch processing pipelines**.

Borde May 2023 - May 2025
MLOps Intern *Remote*

- Designed and operationalized automated **end-to-end data preprocessing and feature engineering pipelines** for 120K+ images using Python, OpenCV, and automated validation scripts.
- **Fine-tuned**, **containerized (Docker)**, and **deployed vision models** across **Azure** and **AWS GPU instances**, establishing automated monitoring, experiment tracking via **MLflow**, and **CI/CD workflows**.
- Drove a **60% overall model accuracy** improvement and a **150% precision lift on critical minority classes**.

Projects

- **Teaching Gemma-3 to Reason** Github:
  - Drove **12% gain in answer accuracy** and **47% surge in reasoning quality** by architecting a 3-stage SFT + GRPO fine-tuning pipeline for **Gemma-3**.
  - **Developed custom composite reward function** with weighted objectives (logical coherence, mathematical precision, strict XML compliance) that directly shapes model behavior at the reward-signal level.
- **Dynamic Model Router for Tau** Github:
  - Architected a **runtime LLM routing engine for Tau coding agent**, dynamically selecting models based on task complexity, context-window requirements, latency, cost, and model capabilities through **configurable routing policies**.
  - Designed a modular **routing layer with session-aware model switching, fallback handling, and observability hooks**, enabling Tau to support configurable multi-model execution without coupling routing logic to the core agent loop.
- **Secure Multi-Tenant RAG System** Github:
  - Engineered an enterprise **RAG pipeline** using **LLaMA 3** and **ChromaDB** vector database, enforcing granular **Role-Based Access Control (RBAC)** via metadata filtering to ensure strict multi-tenant data isolation.
  - Architected high-throughput **FastAPI** microservices integrated with a **Streamlit** frontend and stateful session management, streamlining authenticated cross-department document retrieval.

Education

Rajiv Gandhi Institute of Petroleum Technology 2022 - 2027
Dual Degree (B.Tech in Computer Science & M.Tech in AI)

Achievements

- **Winner Siemens Aavishkar Hackathon 2026**.
- **Meta PyTorch OpenEnv Hackathon** Grand Finalist (Top 100 out of 3500+ teams).
- **Ranked 52 in Amazon ML Challenge 2025** all over India.
- Grand Finalist at **Dark Patterns Buster Hackathon** organised by the Dept. of Consumer Affairs.
- **GATE DA 2026** Qualified AIR 6110.

Skills

- **Programming Languages**: Python, C/C++, Matlab, SQL, Bash
- **Libraries & Frameworks**: PyTorch, TensorFlow, Keras, NumPy, Pandas, Matplotlib, Seaborn, CUDA, Git, Triton, Scikit-learn, OpenCV, Hugging Face, LangChain, LangGraph, LangFuse, LangSmith, Guardrails, MongoDB, PostgreSQL, Vector Databases
- **Cloud & MLOps**: MLflow, DVC, Apache Airflow, Docker, FastAPI, Flask, Weights & Biases (WandB), ChromaDB, Pinecone, Azure ML, AWS, DataBricks, Linux, Postman, Raspberry Pi, Ollama, llama.cpp, vLLM