

TARIFUL ISLAM TARIF

AI Engineer | Machine Learning Engineer | Data Scientist

Dhaka, Bangladesh | +880 1838-051084 | tariful639@gmail.com | [GitHub: imTariful](https://github.com/imTariful) | [LinkedIn: tariful-islam-tarif](https://www.linkedin.com/in/tariful-islam-tarif)

PROFESSIONAL SUMMARY

AI/ML Engineer with hands-on experience designing, deploying, and operating production-grade machine learning, LLM, and data-driven systems. Skilled in building scalable, low-latency inference pipelines, multi-agent LLM architectures, and end-to-end MLOps workflows using Python, FastAPI, Docker, and Kubernetes. Proven ability to translate reliability and product requirements into measurable engineering outcomes, including sub-100ms inference latency, 80%+ cache-hit efficiency gains, and automated LLM evaluation pipelines. Strong foundation in statistics, data mining, and deep learning through a B.Sc. in Data Science & Engineering (CGPA 3.54/4.00). Seeking to apply expertise in machine learning, NLP, and applied AI research to deliver reliable, production-ready AI solutions.

TECHNICAL SKILLS

Programming Languages: Python, SQL, C++, Java, C

Machine Learning & Deep Learning: PyTorch, TensorFlow, Scikit-learn, ONNX Runtime, Model Training & Evaluation, Model Versioning

LLMs & Generative AI: LangChain, LangGraph, Retrieval-Augmented Generation (RAG), Prompt Engineering, Multi-Agent Systems, LLM-as-Judge Evaluation, OpenAI API, Hugging Face

Data Science: Data Mining, Statistical Analysis, Exploratory Data Analysis, NL-to-SQL, Feature Engineering

Backend & APIs: FastAPI, REST API Design, Asynchronous Programming, Microservices Architecture

Databases: PostgreSQL, MongoDB, SQLite, Vector Databases (FAISS, Pinecone, Chroma)

MLOps & Observability: Prometheus, OpenTelemetry, Structured Logging, Drift Monitoring, Model Versioning

DevOps & Tools: Docker, Kubernetes, Git, GitHub, Canary Rollouts, Autoscaling

PROFESSIONAL EXPERIENCE

Forward Deployed Engineer — INFINOZ

March 2026 – Present

infinoz.com | Dhaka, Bangladesh

- Partnered directly with product and client teams as a Forward Deployment Engineer to design, build, and deploy applied AI solutions from research through to production.
- Architected and developed InfiChat, an AI-driven conversational assistant, owning model integration, system architecture, and end-to-end deployment.
- Led R&D on a real-time Voice Agent, integrating speech recognition (ASR), natural language understanding (NLU), and speech synthesis (TTS) into a unified conversational pipeline.
- Fine-tuned and evaluated Bangla Text-to-Speech (TTS) models, covering dataset preparation, model fine-tuning, and voice-quality evaluation to improve native Bengali speech synthesis.
- Collaborated cross-functionally with engineering and product stakeholders to convert experimental AI research into reliable, scalable, production-ready voice and chat-based AI components.

KEY PROJECTS

Intelligence Reliability Platform — LLM Observability & Evaluation

FastAPI, React, OpenAI API, Ollama, SQLAlchemy, Docker | [GitHub: imTariful/LLM-evaluation](https://github.com/imTariful/LLM-evaluation)

- Built a production-grade LLM observability platform featuring prompt versioning, risk controls, and multi-model fallback routing (OpenAI → Ollama → Mock) for high availability.
- Designed automated LLM evaluation pipelines using ensemble LLM-as-judge scoring across quality, safety, and hallucination-detection dimensions.
- Implemented real-time trace logging, cost tracking, latency metrics, and drift monitoring to support continuous model governance.
- Developed role-based access control (RBAC), audit logging, and secure inference workflows for enterprise-grade compliance.

Real-Time ML Inference System

FastAPI, ONNX Runtime, Redis, Prometheus, Docker, Kubernetes | GitHub: [imTariful/Real-Time-ML-Inference-System](#)

- Engineered a low-latency ML inference API supporting multi-model and multi-version serving.
- Achieved sub-100ms inference latency using ONNX Runtime and Redis-backed intelligent caching, cutting redundant compute by 80%+.
- Implemented Prometheus metrics, health checks, and OpenTelemetry-based structured logging for full-stack observability.
- Deployed via Docker Compose and Kubernetes with autoscaling and canary rollout support for zero-downtime releases.

Multi-Agent Medical Data Assistant

Python, Google Gemini, GPT-4o-mini, SQLite, NL-to-SQL | GitHub: [imTariful/multi-agent](#)

- Built a multi-agent LLM system enabling natural-language querying of medical datasets (heart disease, cancer, diabetes).
- Implemented NL-to-SQL pipelines with automated routing between database query tools and medical web-search agents.
- Integrated LLM-based summarization to generate structured, evidence-based analytical insights for healthcare data.

Advanced Image Change Detection System

Python, OpenCV, ORB, SSIM | GitHub: [imTariful/Change-detection](#)

- Designed a computer vision pipeline for robust before/after image comparison under camera misalignment.
- Implemented ORB feature matching, homography alignment, and SSIM-based structural change detection.
- Generated annotated outputs, heatmaps, binary masks, and structured JSON summaries to support auditability.

LangGraph-Style Multi-Agent Debate Workflow

Python, LangGraph, Graphviz | GitHub: [imTariful/agent_debate](#)

- Built a deterministic multi-agent orchestration system with strict turn control and debate sequencing logic.
- Implemented memory nodes, semantic duplicate detection, and replayable JSON-based event logging.
- Visualized multi-agent workflows as DAGs to support debugging and evaluation.

EDUCATION

B.Sc. in Data Science & Engineering2024 – Present

University of Frontier Technology Bangladesh (UFTB)

CGPA: 3.54 / 4.00

Relevant Coursework: Machine Learning, Deep Learning, Natural Language Processing, Data Mining, Statistics, Python Programming

ACHIEVEMENTS & CERTIFICATIONS

- 9th Place, National Datathon
- Finalist, VisionX Competition
- AI Engineer Bootcamp, Ostad
- Active open-source contributor on GitHub ([github.com/imTariful](#))

LANGUAGES

Bangla (Native) | English (Professional Working Proficiency)