MD TANVIR ALAM

AI Engineer | Generative AI & Agentic Systems | CTO, Alawaf Technologies Limited

Dhaka, Bangladesh | +8801716416679 | tanviralamshuvo@gmail.com | cto@alawaf.com.bd

GitHub: github.com/tanvir-ux | Portfolio: tanvir-ux.github.io | LinkedIn: linkedin.com/in/mdtanviralam

Open to: Remote from Bangladesh, or APAC (Singapore EP, Jakarta, India)

SUMMARY

AI engineer and Chief Technology Officer with 5+ years of software engineering experience, focused on generative AI, LLM applications, and agentic systems. Builds Python and FastAPI services for AI agents, RAG and GraphRAG retrieval, Model Context Protocol (MCP) tool servers, LLM evaluation, and LoRA fine-tuning and serving. Backed by production Java, Spring Boot, Kafka, and Kubernetes microservices for regulated banking, including contract customization of Oracle FLEXCUBE and OBPM for Bank Danamon and SMBC. Open-source fixes landed in Meta Infer, Google Gson, and Google Auto.

SKILLS

Generative AI and Agents: LLM applications, AI agents, agentic workflows, multi-agent orchestration, tool calling, human-in-the-loop approval, Model Context Protocol (MCP), FastMCP, A2A (Agent2Agent), prompt engineering

LLM Frameworks and Models: LangChain, LangGraph, LlamaIndex, OpenAI Agents SDK, OpenAI API, Gemini, Anthropic Claude, Ollama, Qwen

Retrieval and Evaluation: RAG, GraphRAG, knowledge graphs, embeddings, hybrid search (BM25 + vector), reciprocal-rank fusion, reranking, FAISS, pgvector, ChromaDB, golden-set evaluation, retrieval benchmarks

ML and Fine-Tuning: PyTorch, Hugging Face Transformers, PEFT, LoRA, TensorFlow, Keras, OpenAI-compatible model serving

Backend and Data: Python, Java, TypeScript, FastAPI, Spring Boot, Microservices, REST API Design, OpenAPI, Node.js, Grails, PySpark, Delta Lake, Databricks, PostgreSQL, Oracle PL/SQL (CDB/PDB), MySQL, Redis, Flyway

Architecture, Cloud and Security: Distributed Systems, System Design, Event-Driven Architecture, DDD, CQRS, AWS (EKS, ECS, Lambda, S3), GCP, Azure, Docker, Kubernetes, Apache Kafka, MQTT, Nginx, GitHub Actions CI/CD, OpenTelemetry, OAuth2, OIDC, JWT, RBAC

EXPERIENCE

CTO / Senior Software Engineer | Alawaf Technologies Limited
November 2026 – Present

- Build generative AI applications, LLM solutions, agentic workflows, RAG pipelines, and knowledge graphs.
- Work on distributed systems and Java / Spring backends used in regulated environments.

Freelance Senior Software Engineer | Various US Clients (Remote)
Nov 2025 – Oct 2026

- Delivered Java / Spring Boot microservices and REST/SOAP integrations for multiple US clients, working remotely across US time zones.
- Implemented SSO and IAM features with OIDC, OAuth2, and SAML2, issuing and validating JWT access tokens through Spring Security.
- Built full-stack features with Angular and Vue.js front ends in TypeScript and Node.js services on top of Java backends.
- Designed Kafka-based event-driven services and containerized deployments on Docker and Kubernetes with CI/CD pipelines.

Lead Software Engineer / AI Consultant | MehvozSoft
Aug 2025 – Oct 2025

- Designed and developed self-executing AI agents with LangChain, orchestrating multi-step reasoning to reduce manual oversight by 40%.
- Engineered RAG architectures so private language models can query internal document repositories and call external APIs via tool-calling.
- Integrated resilient messaging pipelines with Kafka and Redis to streamline background processing and improve Java / Spring Boot performance.

Senior Software Engineer (Java) | Midas Software Solutions Ltd.
Mar 2023 – Aug 2025

- **Bank Danamon (Indonesia):** Architected and executed migration of a core digital banking monolith to a distributed Spring Boot Microservices structure, delivering 70% faster deployment cycles for high-concurrency instances.
- Designed, optimized, and maintained over 100 microservice REST endpoints using OAuth2 and JWT, supporting 50k+ concurrent users.
- Contract engagement: customized Oracle FLEXCUBE and Oracle Banking Payments (OBPM) for Bank Danamon (Indonesia) and SMBC (Singapore) using JavaScript, XML, and Oracle PL/SQL to execute cross-border SWIFT transactions.
- Applied System Design and Distributed Systems practices (Kafka, Kubernetes, Docker, CI/CD) on production banking workloads with PostgreSQL and Oracle.

Application Developer | MehvozSoft
Sep 2021 – Feb 2023

- Developed and deployed cross-platform mobile utilities (including Express Entry Calculator and Adhaan applications) on iOS and Android using Ionic and Angular.
- Crafted database schemas and optimized SQL queries, indexing, and procedures on PostgreSQL and MySQL to accelerate transactional execution.

SELECTED PROJECTS

Original portfolio demos, not employer or client production systems. Repositories: github.com/tanvir-ux

ai-engineering-stack (*demo*) | *Python, FastAPI, LangGraph, OpenAI Agents SDK, FastMCP, PySpark, React*

- Built an enterprise knowledge copilot in which a LangGraph workflow with durable checkpoints runs an OpenAI Agents SDK agent, streams answers over SSE, and pauses for human approval before any write action.
- Carried user identity end to end: OIDC (authorization code + PKCE) and JWT verification at the FastAPI gateway, with role-based access re-checked in the FastMCP tool server and applied before retrieval ranking.
- Implemented hybrid retrieval (BM25, FAISS vectors, reciprocal-rank fusion, reranking) over a PySpark and Delta Lake bronze/silver/gold pipeline with de-duplication, chunking, and embeddings.
- Added a 32-question golden set, retrieval benchmark, access-control leak test, and end-to-end agent eval against a deterministic test model, plus OpenTelemetry tracing; CI builds Docker images and validates a Helm chart.

agent-platform (*demo*) | *Python, FastAPI, LangGraph, Google ADK, AutoGen, CrewAI, MCP, A2A*

- Implemented one customer-support agent task in eight setups (LangGraph, Google ADK, OpenAI Agents SDK, Microsoft Agent Framework, AutoGen, Semantic Kernel, CrewAI, and OpenAI Agents SDK over MCP) behind one runner interface.
- Exposed the shared tools as a FastMCP server (stdio and streamable HTTP) and the agent as an A2A agent with an agent card and JSON-RPC endpoint, served from one FastAPI app with OpenAPI docs.
- Built a deterministic OpenAI-compatible mock LLM so every framework runs in CI without API keys (90 tests), and completed a live Gemini run through the OpenAI Agents SDK.

graphrag-service (*demo*) | *Python, FastAPI, LlamaIndex, LangChain, networkx*

- Compared vector RAG (native, LlamaIndex, and LangChain retrievers) with GraphRAG built from triple extraction, a networkx knowledge graph, Louvain community reports, entity linking, and multi-hop path search.
- Built a 17-question evaluation harness for document recall and answer accuracy; on an author-written synthetic corpus, GraphRAG reached 0.88 accuracy vs 0.47 for vector RAG (multi-hop 1.00 vs 0.43).
- Served the pipelines through FastAPI with a rate-limited optional Gemini mode; Pinecone, Azure OpenAI, and Amazon Bedrock adapters are tested against mocks and stubs only.

llm-finetune-serve (*demo*) | *Python, PyTorch, PEFT, LoRA, Hugging Face Transformers, FastAPI*

- Fine-tuned SmolLM2-135M-Instruct with LoRA in a plain PyTorch and PEFT loop on CPU (3.5% of parameters trainable, loss masked to the answer) to return strict JSON for support-ticket triage on synthetic data.
- Evaluated base vs tuned on 100 held-out synthetic tickets: valid JSON rose from 0% to 100%, with 90% exact match on seen-style tickets and 35% on unseen wording.
- Served the adapter behind an OpenAI-compatible FastAPI endpoint called with the OpenAI SDK, merged adapter weights for deployment, and covered the pipeline with 23 offline tests in GitHub Actions.

agentic-mcp-gateway | *Python, FastAPI, MCP, RAG, Docker.*

MCP server plus FastAPI HTTP agent sharing one tool registry for document retrieval, payment searches, and ticket creation. ReAct loop with CI stubbing and configurable live LLM providers. The default path runs without an API key.

banking-payments-spring | *Java 17, Spring Boot 3.3, Kafka, PostgreSQL, Docker.*

Credit-transfer REST microservice with JWT resource security, idempotent POST, JPA persistence, an AML sanctions stub, and ISO 20022 / SWIFT-inspired Kafka domain events.

kafka-spring-microservices | *Java, Spring Boot, Kafka, Kubernetes, PostgreSQL.*

Transfer, ledger, and notifications services with double-entry execution, idempotent consumers, publish-after-commit events, plus Docker Compose and Kubernetes manifests.

OPEN SOURCE

- **Meta Infer:** Track streams from Socket.getInputStream and getOutputStream in Pulse (PR #2302).
- **Meta Infer:** Track Deflater and Inflater as native resources in Pulse (PR #2275).
- **Google Gson:** Record the JSON path after promoteNameToValue (PR #3098).
- **Google Auto:** Copy javax and jakarta @Qualifier onto generated AutoFactory parameters (commit d10c050).

EDUCATION

Stamford University Bangladesh — B.Sc. in Electrical and Electronic Engineering (EEE) *Graduated Aug 2021 | CGPA: 3.25*

Machine Learning Specialization — Stanford University and DeepLearning.AI (Coursera) *Credential ID: 4UGJY8KESNCZ*