Ahammad Nafiz

Based in Dhaka, Bangladesh

+880 1619960303
ahammadnafiz.github.io
github.com/ahammadnafiz
linkedin.com/in/ahammad-nafiz
ahammadnafiz86@gmail.com

Data Science student and ML engineer working on RAG systems, LLM agents, and mechanistic interpretability. Aiming for graduate research in Europe and a role at a frontier AI lab.

EXPERIENCE

Markopolo AI — *Associate Product Engineer, DeepDive*
August 2026 - Present
– Building DeepDive, an AI social listening platform, plus an AEO visibility tool and an AI data analysis agent, across a TypeScript and Python LLM stack.

AskTuring.ai — *Jr. Software Engineer*
September 2025 – May 2026
– Owned the Eval Service and admin dashboard, automating RAG benchmark evaluation across HotpotQA, MuSiQue, FinQA, and CUAD.
– Architected Deepen Answer, a production RAG pipeline combining extended-thinking LLMs, an agentic gap-fill layer, and hybrid BM25 + vector search with reciprocal rank fusion.
– Built FastAPI backend services that improved retrieval accuracy and reduced query latency.

United International University — *Research Assistant*
December 2025 – February 2026
– Research autonomous AI agents and multi-agent systems, contributing to publications on LLM applications, information retrieval, and agentic workflow design.

PROJECTS

Atlas — *Source Control for Coding Agents*
github.com/tryatlas/atlas
– Core contributor to an open-source agentic IDE (Rust, Tauri, React, 9.4k+ stars) that links commits to LLM agent sessions.
– Built Team Timeline, Org Agent, and Atlas Memory, a multi-agent memory layer with HNSW vector search.

Buyziz — *AI Commerce Platform*
buyziz.com
– Built the AI sales agent (multimodal, Bangla/English, tool calling) and the store builder, and led product engineering for a multi-tenant e-commerce SaaS.

fizzlabai — *Deep Research Agent*
github.com/ahammadnafiz/fizzlabai
– Built a solo deep research agent that runs a deterministic 8-phase pipeline (plan, search, read, synth, verify, re-plan, polish, close) over the open web, producing cited Markdown answers with a full execution trace.
– Implemented per-phase LLM routing across Gemini Flash tiers on FastAPI, LangGraph, and Postgres/pgvector, with OpenTelemetry tracing into Grafana, Tempo, and Loki.

Engram — *Persistent Memory Infrastructure for AI Agents*
github.com/ahammadnafiz/engram
– Built a memory library for AI agents on PostgreSQL and pgvector, combining hybrid search (BM25 + vector + RRF), graph traversal, and durable task ledgers.
– Cut LLM context cost by up to 15x while holding over 90% accuracy on LongMemEval, LoCoMo, and BEAM long-term memory benchmarks.

Personal Knowledge Assistant — *Multi-Agent RAG System*
github.com/ahammadnafiz/Personal-Knowledge-Assistant
– Built a multi-agent RAG system with confidence-based routing, knowledge-strip decomposition, and automatic web-search fallback for low-confidence queries.

FizTorch — *Lightweight Deep Learning Framework*
github.com/ahammadnafiz/FizTorch
– Built a NumPy-based deep learning framework from scratch with an autograd engine, core layers, and an Adam optimizer, validated on MNIST and California Housing.

EDUCATION

United International University, Dhaka — *B.Sc. in Data Science*
September 2023 – Present
– Machine Learning Specialization & Deep Learning Specialization — DeepLearning.AI and Stanford University (Coursera)

SKILLS

Languages — Python, C++, SQL, Bash, CUDA

ML / DL — PyTorch, scikit-learn, XGBoost, CatBoost, NumPy, Pandas, OpenCV, Matplotlib

LLM & GenAI — LangChain, LlamaIndex, Hugging Face Transformers, CrewAI, AutoGen, LangGraph, vLLM, OpenAI API, Anthropic API

AI Agent Stack — ReAct, Tool Use, Function Calling, Multi-Agent Systems, Supervisor/Swarm/Hierarchical Architectures, LLM Memory (Episodic, Semantic, Working), RAG, LLM Eval, Agentic Pipelines

Fine-tuning — LoRA, QLoRA, PEFT, SFT, RLHF, Prompt Engineering, Instruction Tuning

Inference — vLLM, ONNX, TensorRT, BentoML, TorchServe, Quantization

Backend — FastAPI, Streamlit, REST APIs, Docker, Linux, Kubernetes

MLOps — MLflow, Weights & Biases, DVC, GitHub Actions, CI/CD, Git, Jupyter

Databases — MySQL, PostgreSQL, pgvector, Chroma, FAISS, Pinecone, Weaviate

Domains — RAG, LLM Evaluation, AI Agents, NLP, Computer Vision, MLOps, GenAI, Mechanistic Interpretability

ACHIEVEMENTS

Champion, IC6 University Innovation Hub Program — UIU Cohort
Oct 2025
Earned BDT 60,000 in pre-seed funding after a 15-week program. Issued by University Innovation Hub Program.

Best Prototype Award — Blockchain Olympiad, UIU CSE Fest
Jan 2025
Team Auctra won Best Prototype for a blockchain-based solution. BDT 20,000 prize. Issued by United International University.

Champion, CSE Project Show 2024 (Spring) — OOP for Data Science
Jun 2024
Predicta — an end-to-end data analysis and ML platform for EDA, preprocessing, and modeling. Issued by United International University.

1st Runners Up, Data Innovators Challenge 2024
Sales analytics track.