Alif Tasbir

aliftasbir@gmail.com | Leicester, UK | github.com/a1if | a1if.github.io

Right to work in the UK: full-time now; applying for the Graduate visa in December 2026; no sponsorship needed to start

PROFILE

MSc Artificial Intelligence graduate of De Montfort University who designs, fine-tunes and evaluates machine learning and LLM systems in Python and PyTorch: LoRA fine-tuning, hybrid-retrieval RAG, multi-agent workflows and MCP tool servers. Dissertation work measures accuracy, throughput and energy use on constrained hardware, alongside three peer-reviewed papers in clinical AI. Seeking a Graduate AI or Machine Learning Engineer role.

EDUCATION

MSc Artificial Intelligence, De Montfort University, Leicester, UK | Sep 2025 to Sep 2026
Grade: Distinction (80%)
Modules: Neural Networks and NLP; Research Methods and Applied AI; AI for Mobile Robots; Fuzzy Logic and Evolutionary Computing.

BSc Computer Science and Engineering, University of Information Technology and Sciences (UITS), Dhaka, Bangladesh | Aug 2018 to Aug 2022
CGPA 3.74/4.00. Modules: Machine Learning, Data Structures and Algorithms, Software Engineering, Object Oriented Programming.

TECHNICAL SKILLS

- **LLM systems:** RAG (hybrid search, reranking, multi-hop retrieval), LangGraph agents, Model Context Protocol (MCP), tool use, Ollama, Gemini API, Hugging Face Transformers
- **Fine-tuning and evaluation:** LoRA, QLoRA, PEFT, TRL, LLM evaluation (EM/F1, faithfulness, significance testing, BLEU, ROUGE, BERTScore), MLflow experiment tracking
- **Retrieval:** FAISS, ChromaDB, BM25, Sentence Transformers, cross-encoder reranking
- **Languages and engineering:** Python, C, C++, Java, JavaScript; object-oriented design; Git, GitHub Actions (CI), Docker, DVC, OpenTelemetry
- **APIs and apps:** FastAPI, Flask, Django, Streamlit, Gradio
- **ML and data:** PyTorch, TensorFlow, Keras, scikit-learn, Pandas, NumPy, SQL (MySQL), NoSQL

SELECTED PROJECTS

Fine-Tuned Small LLM for Multi-Hop RAG | Python, PyTorch, PEFT, FAISS, LangGraph, MCP, FastAPI, OpenTelemetry
- Fine-tuned Qwen3-1.7B with LoRA for multi-hop question answering on HotpotQA, raising exact match from 10.4 to 58.4 and matching a 4.7× larger Qwen3-8B (71.2 vs 69.3 F1) with ~5× smaller weights at ~8× the throughput.
- Built hybrid retrieval (BM25 + FAISS dense search, rank fusion, cross-encoder reranking) with iterative multi-hop search, raising evidence recall from 54% to 89%; trained the model to cite sources and abstain without evidence.
- Exposed it as an MCP server with typed tools and OpenTelemetry tracing, fine-tuned a LangGraph agent for tool use (+19.7 F1), and made training 2.8× faster with 55% less GPU memory.

Multimodal RAG Application | Python, FastAPI, Streamlit, FAISS, BM25, Tesseract OCR, Gemini, Groq
- Built a bilingual (Bangla and English) RAG chatbot over PDFs, scans, Office files and images, with OCR and page-level citations.
- Implemented hybrid retrieval: BM25 and multilingual-e5 embeddings in FAISS, merged with reciprocal rank fusion.
- Added LLM-routed web search that decides when fresh data is needed and rewrites the query; streams Gemini or Groq answers.

Heterogeneous Multi-Agent Edge AI for Clinical Decision Support (MSc dissertation)
- Built a two-agent ECG pipeline: a CNN-LSTM heartbeat classifier sends events to a frozen Gemma 4 E4B model, either as compact text or through a trained linear adapter that emits 4 virtual tokens per event, comparing two ways of constructing the model's context.
- Cut the language model's prompt-processing time by 24–78% against compact text, and improved balanced accuracy at 10 and 20 events (0.83 vs 0.73), with held-out patients, three training seeds and patient-level bootstrap statistics.
- Ran batched serving on a 12 GB GPU: 10–30% higher median throughput and 12–28% lower median energy per decision than filtered text. Measured energy with NVML power sampling at 100 Hz.

WORK EXPERIENCE

Junior Filing Specialist, Compliance Services Limited (part of WSD Group), Dhaka, Bangladesh | Apr 2024 to Sep 2025
- Handled live filings and manually proofread documents to verify accuracy.
- Converted Word documents to HTML and built webpages using HTML and CSS.
- Maintained filing records in Excel, kept documentation up to date and analysed data.

RESEARCH EXPERIENCE

Research Assistant, University of Information Technology and Sciences (UITS), Dhaka, Bangladesh | Jul 2022 to Mar 2023
- Developed graph-based deep learning models for epileptic seizure detection from EEG signals, improving detection accuracy.
- Applied signal processing, CNNs and statistical inference to optimise model results.
- Contributed to grant proposals and experiment design.

PUBLICATIONS

1. Parkinson's disease detection from EEG signal employing autoencoder and RBFNN-based hybrid deep learning framework utilizing power spectral density. *Digital Health*, SAGE, 2024.
2. Graph attention network and radial basis function neural network-based hybrid framework for epileptic seizure detection from EEG signal. *Journal of Autonomous Intelligence*, Vol. 7, Issue 3, 2023.
3. Epileptic seizure detection from EEG signals using linear graph convolutional network and DenseNet based hybrid framework. *Journal of Radiation Research and Applied Sciences*, Elsevier, Vol. 16, 2023.