Abesh Chakraborty

AI/ML Engineer – Kolkata, India
abeshchakraborty10@gmail.com LinkedIn
GitHub Leetcode

Education

Institute of Engineering and Management (IEM), Kolkata 2024 - 2028
(School of University of Engineering and Management, Kolkata)
B.Tech in Computer Science and Engineering (Artificial Intelligence) *CGPA: 8.79*

Relevant Coursework: Data Structures & Algorithms, Object Oriented Programming, Artificial Intelligence & Machine Learning, Database Management Systems, Operating Systems.

Technical Skills

- **Languages & Tools:** Python, C/C++, SQL, Linux, Docker, Git/GitHub, FastAPI, Streamlit, Jupyter
- **AI/ML Frameworks:** PyTorch, Hugging Face, LangChain, LangGraph, Scikit-learn, Pandas, NumPy
- **Domains & Concepts:** GenAI (LLMs, RAG, Agents, Fine-Tuning), NLP, Deep Learning, Computer Vision

Experience

AI Generalist Intern – Gaudi Engineer Pvt. Ltd. May 2026 - Present
Python, FastAPI, MongoDB, Docker, Anthropic Agent SDK, MCP

- Contained untrusted LLM-generated shell commands with a 3-tier command guard over an OS-level filesystem and network **sandbox**, covered by 30 tests asserting dangerous commands block while ordinary work still runs.
- Raised top-5 retrieval accuracy from 88% to 93% at 0.3 s/query on CPU across 5,786 standards clauses, fusing **BM25** and **BGE** search with **RRF** and a cross-encoder reranker.
- Removed per-model integration work across 5 applications by deploying **MCP** servers that standardized tool execution and database connectivity.

Projects

qwen.cpp: CPU Inference Engine for Qwen2.5-0.5B *Aug 2026 - Sept 2026*
C++17, Safetensors, Byte-Level BPE, Grouped-Query Attention, RoPE

- Ran Qwen2.5-0.5B end to end at 4 s/token by implementing grouped-query attention (14 query / 2 KV heads), RoPE, RMSNorm, and SwiGLU in C++17 with no BLAS or ML framework.
- Decoded a 151,936-token vocabulary by hand-writing a byte-level BPE tokenizer with ChatML templating, and parsed 988 MB of bf16 safetensors into fp32 without a deserialization library.

Argus: Real-Time Traffic Monitoring & Helmet Violation Platform *Feb 2026 - Present*
Python, PyTorch, RT-DETR, ByteTrack, FastAPI, SQLite, ONNX, OpenCV

- Detected 14 vehicle classes and helmet compliance at 95.3% mAP50 and 18 FPS, a 1.3× speedup, by training NMS-free RT-DETR models and exporting to ONNX.
- Cut false violations from a weak detector (0.489 mAP50-95) by replacing single-frame calls with a 15-frame majority vote over ByteTrack-associated vehicle tracks.
- Turned raw detections into flow, occupancy, and direction analytics via a 10-endpoint FastAPI service with live streaming and SQLite persistence.

ChronoFlow: Real-Time Traffic Analytics on a Time-Series Pipeline *May 2026 - Sept 2026*
Python, FastAPI, SQLModel, TimescaleDB, Docker

- Sustained 181K rows/sec bulk ingest on a TimescaleDB hypertable, load-tested to 27M events across 31 daily chunks with 3-month auto-retention.
- Cut live rollup latency from 4.2 s to 20 ms at 27M rows by adding a since filter so the query prunes chunks instead of scanning the full hypertable.

Certifications

- DeepLearning.AI's Machine Learning Specialization by Andrew Ng (Coursera) *Sept 2025*
- Introduction to Machine Learning - IITKGP (NPTEL) *July 2025*