Md. Saidur Rahman (Pulok)

Dhaka, Bangladesh saidur.pulok@gmail.com +880 1621 299536
linkedin.com/in/saidurpulok github.com/saidurpulok saidurpulok.github.io

Summary

Machine Learning Engineer shipping production ML and LLM systems: LLM evaluation and fine-tuning, model routing, RAG, and computer vision, plus the pipelines and infrastructure behind them (Argo, Airflow, Kubernetes / EKS, AWS, Snowflake). Independent work on LLM serving, routing, and inference-time reasoning.

Experience

Machine Learning Engineer I, Optimizely **Apr 2026–Present**

- Designed a 46-task, 13-category benchmark (medium / hard) to evaluate Optimizely's AI desktop app, Mark (formerly Opal), against third-party AI tools across real work tasks.
- Built an on-device LiteLLM routing proxy whose decision model sends each request to a model sized for the task, cutting spend from defaulting to expensive models; used daily by the team.
- Designed an Argo Workflows ML pipeline on EKS and productionized 2 ML pipelines end to end on Snowflake data; experiments tracked in MLflow and SageMaker.
- Ran QLoRA fine-tuning R&D on Qwen2.5 (0.5B / 1.5B) for customer-data recall: good on facts, weak on numbers.

Machine Learning Engineer, ACI PLC **Aug 2025–Apr 2026**

- Productionized and deployed an on-premise face-recognition attendance system replacing fingerprint check-in for 50+ employees: FastAPI and Redis Streams pipeline over RTSP cameras, InsightFace on ONNX / CUDA, Qdrant.
- Built and launched AgroGPT, a RAG platform for precision agriculture used by farmers, agri-businesses, and researchers: Airflow pipelines for weather and market data, FastAPI services, Prometheus / Grafana monitoring.

Jr. Software Engineer, Brick Line Technology Ltd. **Mar 2025–Jul 2025**

- Built Python backend services for Zeda, a US clinical trial management system (CTMS) used by several US clinics: FastAPI, AWS RDS, DynamoDB, S3, Lambda.
- Implemented MFA with AWS Cognito and Twilio, and shipped several Generative AI features.

ML Systems Projects

AxiomServe: LLM Serving Runtime from Scratch **2026–Present**

- Built an LLM serving runtime with batched prefill and manual decode (no `model.generate()`), KV-cache reuse, continuous batching, admission control, FCFS / shortest-prompt-first scheduling, streaming, and cancellation.
- Continuous batching cut p95 time-to-first-token to 119 ms, versus 503 ms for static batching and 3.48 s sequential, at 8 concurrent requests (Qwen2.5-0.5B, RTX 5070 Laptop GPU); static batching kept the highest throughput.

routewise: Cost–Quality Benchmark for LLM Routing **2026**

- Benchmarked learned LLM routers on HELM results (8.5k prompts, 22 models), cross-validated with bootstrap CIs.
- At equal accuracy, routing cost 48% less than the best cheap model (DeepSeek-V3) on 2024 models and 25% less than the best model (Qwen3) on 2025 models, and beat every single model by 3–4.5 points.

VeriSearch: Verifier-Guided LLM Reasoning Under Token Budgets **2026–Present**

- On GSM8K with Qwen2.5-Math-1.5B, early stopping used 35% fewer tokens than full verifier reranking at about 1 point lower accuracy; caught the verifier relying on answer-format shortcuts.

Technical Skills

Languages:	Python, C/C++, TypeScript, SQL
LLM / ML:	PyTorch, Hugging Face (Transformers, PEFT / QLoRA, TRL), vLLM, TensorRT, Triton, bitsandbytes, ONNX / CUDA, LangChain, LiteLLM, OpenCV
MLOps / Data:	MLflow, Airflow, SageMaker, Snowflake, Kafka, PostgreSQL, Redis, Qdrant
Backend / Infra:	FastAPI, AWS, GCP, Kubernetes / EKS, Docker, Argo Workflows, Jenkins, pytest, Prometheus / Grafana

Education

Bangladesh University of Engineering and Technology (BUET) **2019–2024**
B.Sc. in Mechanical Engineering *Dhaka, Bangladesh*

- **Paper:** "Integrating Machine Learning for Enhanced Detection and Localization of Gas Pipeline Leaks," *ICCET* 2024.