

Nazmus Sakib

+880-1759948779 | sakibbuet2k19@gmail.com | Dhaka, Bangladesh

Google Scholar | Portfolio | GitHub | LinkedIn

EDUCATION

Bangladesh University of Engineering and Technology (BUET)

Bachelor of Science in Computer Science and Engineering; CGPA: 3.82/4.00

Dhaka, Bangladesh

Jan 2020 – Mar 2025

- Academic Rank: Top 20% of class based on academic performance.
- Awards: Enlisted thrice in the Dean's List for academic excellence.

EXPERIENCE

BRAC University

Lecturer

Dhaka, Bangladesh

Present

- Instructing undergraduate courses including **Machine Learning**, **Operating Systems**, **Software Engineering**, and **System Analysis & Design**.
- Mentoring students on architectural patterns and bridging the gap between theoretical concepts and industry application.

Sluqe

Software Engineer (Full Stack & AI)

Warszawa, Poland (Remote)

May 2025 – May 2026

- **Sluqe Platform:** Built an AI voice-note platform on **Next.js** and **Firebase Cloud Functions** where recordings are transcribed, structured, and vector-indexed into a searchable corpus; shipped **Genkit** agent flows with in-chat web search and deep research over **Groq**, a **Flutter** iOS/Android client, and **Stripe** tiered billing with server-enforced usage quotas.
- **SOS Alert System:** Engineered a device-agnostic emergency pipeline on **AWS** (API Gateway, Lambda, DynamoDB, SNS to **Twilio** voice calls), provisioned with **Terraform** and tested end-to-end against **LocalStack**; wrote an async TCP bridge with protocol adapters for Tuya WiFi buttons, 4G terminals, and GPS bracelets, reverse-engineering undocumented binary and text framing.
- **Aviators Spring:** Developed a **Flutter** pilot navigation app over a **Mapbox** base map, stacking FAA NASR airport data (a GeoJSON pipeline joining six source tables into simplified and full tiers), **NOAA** METAR/TAF weather, and 3.5 GB of VFR sectional charts published through an automated parallel tileset upload pipeline.

Oleyn

Founding Engineer (Backend & AI)

Dhaka, Bangladesh (Remote)

May 2024 – Dec 2024

- **Shrobon:** Engineered an AI-powered CRM and marketing campaign system featuring RAG chatbots. Developed a Bengali ASR system with fine-tuned models and **NVIDIA NeMo** speaker diarization for customer center analytics.
- **Superattorney:** Architected a legal research assistant using **FastAPI** and **Redis** caching for high-speed document parsing and retrieval.

Priyo

Software Engineer (Backend)

OnSite

Jun 2024 – Dec 2024

- **Priyo-Penguin:** Deployed an AI customer support bot using **Django**, **Nginx**, and **OpenAI**, handling concurrent real-time queries.
- Developed **Dolphin-Priyo**, a warehousing application managing cross-platform users with **REST APIs**.
- Implemented bot protection using **Cloudflare Turnstile** and visualized data via **Metabase**.
- Managed dynamic email campaigns using **SendGrid** and Twilio integration.

Yobo

Software Engineer Intern

Ontario, Canada (Remote)

Jan 2024 – May 2024

- Built **YoboSQL**, a text-to-SQL AI chat interface using **LangChain**, **FastAPI**, and **SvelteKit**.
- Optimized database interactions with MariaDB and deployed on cPanel environments.

TECHNICAL SKILLS

Languages: Python, C++, Java, JavaScript, SQL

Frontend & Mobile: Next.js, React, SvelteKit, Flutter, Tailwind CSS

Backend: FastAPI, Django, Express, Firebase Cloud Functions, AWS Lambda, Spring Boot

AI & Machine Learning: PyTorch, Unsloth, ROCm, vLLM, Supervised Fine tuning, Reinforcement Learning, llama.cpp, LangChain, LlamaIndex, RAG, Genkit, NVIDIA NeMo, HuggingFace

Databases & Vector DBs: PostgreSQL, MySQL, Redis, Firestore, ChromaDB, Qdrant, Supabase, Turso

DevOps & Tools: Docker, AWS (API Gateway, DynamoDB, SNS, SES), Localstack, k8s, Terraform, GCP, Vercel, Nginx, Keycloak, PostHog, Stripe, Git

KEY PROJECTS →

DC-Ops | *GRPO, OpenEnv, vLLM, Unsloth, PyTorch, ROCm*

- Built a physics-grounded reinforcement learning environment for datacenter operations using **OpenEnv**, featuring RC thermal networks and continuous multi-objective reward functions.
- Architected a highly efficient training pipeline on an **AMD MI300X**, generating expert reasoning trajectories with a 32B teacher via **vLLM** for initial Supervised Fine-Tuning (SFT).
- Designed a multi-signal GRPO reward system (physics simulation, scenario-aware heuristics, anti-looping, format compliance) that drove a **+188% composite reward improvement** and a **10× per-step gain** on hard multi-fault datacenter scenarios vs. the base model.

Scout | *FastAPI, Langfuse, Docker, Terraform, Web-Scraping, Next.js, PostgreSQL, Agentic workflow*

- Engineered an autonomous competitive intelligence platform with scheduled agents to detect pricing changes and resolve them into a temporal knowledge graph.
- Architected an exactly-once job queue in **PostgreSQL** and orchestrated hosted browsers to safely render untrusted competitor content for automated weekly briefs.

Medical CoT Assistant | *Unsloth, Llama.cpp, Transformers, LoRA*

- Fine-tuned a 20B parameter LLM on medical datasets to enable clinical Chain-of-Thought (CoT) reasoning.
- Implemented an efficient training pipeline using **Unsloth** and **LLM-as-a-Judge** evaluation; optimized for deployment via **INT4 quantization** and GGUF export.

InsightAI | *Spring Boot, FastAPI, Qdrant, LlamaIndex*

- Architected a microservices-based EdTech platform combining **Study Companion**, **Project Planner**, and **ShopGenie** modules.
- Developed **ShopGenie**, a multimodal e-commerce engine using **CLIP** embeddings for image search and **LLaVA** for product data augmentation.
- Built a RAG pipeline for the **Study Companion** that generates quizzes and flashcards from PDFs and YouTube videos using **LangChain** and **OpenAI**.
- Implemented a unified **Spring Boot** backend for data management and a **FastAPI** service for handling high-concurrency AI inference.

BD Prescription OCR | *PyTorch, ResNet50, CRNN, SWA, MLflow, Grad-CAM*

- Designed a **ResNet50-CRNN** architecture that intercepts backbone feature maps and applies a **bidirectional LSTM** over the width axis, preserving left-to-right structure in handwritten word images for **78-class** medicine-name classification on a Bangladeshi prescription dataset.
- Built a robust training pipeline with **CLAHE** preprocessing, hash-based cross-split leakage detection, staged BatchNorm unfreezing, and **Stochastic Weight Averaging (SWA)** - progressively improving single-model accuracy from 86.08% to 89.82%.
- Ensembled 5 **StratifiedGroupKFold** SWA checkpoints (grouped by image hash to eliminate duplicate leakage), achieving a final test accuracy of **92.53%** and macro-F1 of **0.9234** - a **+6.44 pt** gain over the ResNet50 baseline.

HONORS & ACHIEVEMENTS →

Champion, IUT 11th National ICT Fest Hackathon (2024) – *1st Place in OpenAPI Category*

Champion, CodeCrafters Dev Sprint Hackathon, BUET (2024)

Champion, SUST CSE Carnival Hackathon (2024)

Runners Up, Gen-Dev Hackathon by Acme AI (2024)

Champion, Cefalo ITverse Project Showcase (2023)

Honorable Mention, BCOLBD (Blockchain Olympiad) & IDSOL World Finalist (2024)

7th Place, SEC Inter University Junior Programming Contest (2022)

CERTIFICATIONS

Machine Learning Specialization | *Stanford Online & DeepLearning.AI*

Deep Learning Specialization | *DeepLearning.AI (Andrew Ng)*

PyTorch for Deep Learning Professional Certificate | *DeepLearning.AI*

Intermediate Problem Solver | *Hackerrank*

RESEARCH

Memagent: Developed an AI agent framework for generalized tasks requiring minimal user input. Published in CEUR.

OPEN SOURCE MODELS & DATASETS

Hugging Face Profile (@Melikshah) | *LLM Fine-tuning, NLP Datasets, Multimodal Datasets*

- **GPT-OSS-20B-Clinical-CoT (GGUF)**: Published a 4-bit quantized version of a 20B parameter model fine-tuned for clinical Chain-of-Thought reasoning, optimized for local inference.
- **Multimodal E-Commerce Datasets**: Scraped, curated and released the **Shajgoj** and **General Products** datasets, containing 20,000+ items for image-text retrieval tasks.
- **qwen3.5-4b-base-blindspots**: The Blindspot Discovery Dataset is a highly curated, adversarial evaluation set designed to stress-test and expose the architectural and logical limitations of the Qwen/Qwen3.5-4B-Base model. tasks.