

ANINDYA MAJUMDER

📞 (+880) 1615-791025 ✉ anindyamajumder.54@gmail.com 🌐 anindya.pro 🏠 GitHub 🔗 LinkedIn

EDUCATION	BRAC University Bachelor of Science in Computer Science <i>Dhaka, Bangladesh</i> <i>Oct 2021 – Jun 2025</i> › CGPA: 3.80 out of 4.00 › VC List ×2 & Dean's List ×5 for Academic Excellence › Academic Merit-based Scholarship › Thesis: <i>Hierarchical Transformer-Based Semantic Segmentation of Intraoperative Anatomical Structures</i> — improved Dice scores by up to 28% on challenging organ classes using a DINOv2 backbone with multi-scale cross-attention decoder.
EXPERIENCE	Adjunct Lecturer BRAC University <i>Aug 2026 – Present</i> › Deliver lectures on Data Structures and Algorithms, teaching around 200 undergraduate students per semester. › Grade assignments and examinations, supervise lab sessions, and hold consultation hours to support students with course material and exam preparation. AI Workflow Automation Engineer Riseup Labs Ltd <i>Apr 2026 – Present</i> › Developing LLM-powered automation pipelines for marketing, sales, lead generation, and project management teams, reducing manual operational overhead. › Building role-based multi-agent RAG system over the company's central database, enabling access-controlled, context-aware query responses tailored to each user's role. › Deployed and monitoring hybrid AI inference architecture with locally served open-source models via vLLM for sensitive workloads and cloud APIs for general tasks, optimizing cost, latency, and data security. › Researching RLHF, DPO, and GRPO techniques to improve behavioural alignment, reasoning, safety, and performance optimization of in-house generative AI models. Junior AI Developer Join Venture AI, Betopia Group <i>Oct 2025 – Feb 2026</i> › Integrated Natural Language Query (NLQ) support enabling AI agents to interact with PostgreSQL databases directly, abstracting SQL complexity from end users. › Reduced LLM inference costs significantly by implementing prompt caching, semantic re-ranking, and dynamic model routing in production environments. › Fine-tuned and distilled large generative models into efficient, task-specific variants with lower latency and reduced resource consumption. › Developed multi-agent ed-tech platform with progress tracking and custom MCP servers for enabling seamless agent interaction with internal databases and external APIs. Trainee AI Developer Join Venture AI, Betopia Group <i>Jul 2025 – Sep 2025</i> › Built multi-agent RAG system for a SaaS restaurant receptionist platform, handling end-to-end phone-based order and reservation workflows autonomously. › Developed Graph-RAG pipeline with a re-ranker for a legal chatbot, integrating graph-traversal context with PDF-extracted local retrieval for improved Q&A precision. › Built agentic RAG-based meal and workout planner that generates personalised recommendations from user dietary data and health preferences. › Contributed to mental health conversational agent, with focus on long-context retention and safe, grounded response generation. Undergraduate Teaching Assistant CSE, BRAC University <i>Jun 2024 – May 2025</i> › Tutored Data Structure and Object Oriented Programming to over 100+ students. › Provide consultation, marked quizzes and assignments, and assisted in lab sessions.

TECHNICAL SKILLS	Languages: Python, Java, Go, C++, C#, Bash		
	AI/ML: PyTorch, Scikit-learn, LangChain, LangGraph, vLLM, RAG, Graph-RAG, n8n		
	Backend: Spring Boot, FastAPI, MySQL, Postgresql, MongoDB, Pinecone, Neo4j, Ragas		
	DevOps: Docker, Kubernetes, AWS, Git, GitHub Actions, Nginx, VPS		
PROJECTS	AI Lawyer Graph-RAG Based Legal Q&A Solution		
	› Architected a GraphRAG pipeline with LangChain, Neo4j, Cypher, FastAPI, and LangGraph, combining entity linking, subgraph expansion, hybrid retrieval, and reranking for citation-grounded response generation over a legal knowledge graph. › Established evaluation and observability with RAGAS, tracking faithfulness, relevance, F1 score, and retrieval quality to harden response reliability on ambiguous queries. › Optimised local inference using GGUF/4-bit quantisation, KV cache, FlashAttention, and speculative decoding, achieving faster generation with reduced memory footprint.		
	TrustTaste.de Germany-Based Restaurant SaaS		
	› Built AI voice agents with intent routing, tool-calling, and stateful dialogue for fully automated reservations, ordering, and customer support over phone. › Integrated Twilio and Vapi for speech-to-speech pipelines, implementing event-driven call routing, spam filtering, and human escalation with async callback automation. › Developed OCR-based menu ingestion using vision-language models for structured extraction across multi-modal restaurant data pipelines.		
	Colwords Vocabulary Learning Platform		
	› Developed scalable FastAPI backend with microservice architecture, dependency injection, and layered design for maintainability and independent service scaling. › Built AI-powered voice assistant using ElevenLabs Client SDK to guide users, teach vocabulary, and evaluate learning through AI-generated quizzes. › Optimised performance through Redis caching, rate limiting, memoisation, and database query tuning, reducing response latency across high-traffic endpoints. › Containerised with Docker and deployed via CI/CD pipelines on GitHub Actions to VPS, with logging and distributed monitoring for observability and incident response.		
	NAIS Journal App Micro-blogging Social Media		
	› Built scalable backend with Spring Boot applying layered and domain-driven design for modularity and separation of concerns. › Architected stateless JWT authentication with RBAC, enabling horizontally scalable session management and secure access control. › Optimised MongoDB persistence layer with indexing, connection pooling, and pagination to support high-throughput ingestion and retrieval. › Engineered real-time notifications via Server-Sent Events (SSE), decoupling read-heavy streams from core paths to improve overall responsiveness.		
AWARDS	› Finalist , University Rover Challenge 2024, Utah, US		2024
	› Ranked 39th , Machine Learning Olympiad 2024		2024
	› Ranked 2nd in Asia , University Rover Challenge 2023, Utah, US		2023
	› Finalist , IEEE CS Spectrum: Code Crash 2023		2023
CERTIFICATIONS	› Containerization and Virtualization Skill-track, DataCamp		2026
	› Developing Applications with LangChain Skill-track, DataCamp		2025
	› Machine Learning Specialization, Stanford University & DeepLearning.AI		2024