Md. Motahar Mahtab

+1 226-758-7661 | md.motahar.mahtab@g.bracu.ac.bd | Linkedin | GitHub | Portfolio

Summary

AI Engineer and Applied Researcher with 3+ years of experience designing, training, and deploying intelligent systems across **NLP, Computer Vision, Reinforcement Learning, and Multimodal AI**. With a strong academic foundation (**CGPA 3.99**) and publications in top-tier venues (**NAACL, EMNLP**), I specialize in bridging research innovation with scalable engineering. My work spans building **multi-agent LLM pipelines**, **SOTA NLP models**, and **vision-driven information extraction systems**, along with architecting high-performance MLOps infrastructures leveraging **TensorRT, Triton, Kubernetes, and KEDA**. Passionate about developing efficient, reliable, and explainable AI systems that generalize across domains and modalities.

Education

BRAC University — Dhaka, Bangladesh
CGPA: 3.99; B.Sc in Computer Science & Engineering — *May 2018 – May 2022*

Experience

Sr AI Engineer, Co-Lead (Remote) — Oct. 2024 – Present
Delineate Inc. — *Cambridge, MA, USA*

- **Developed Multi-Agent Data Extraction Pipeline** for diverse files (PDF, DOCX, Code), **outperforming** general-purpose LLMs (Gemini 2.5 Pro, Claude Opus) by **15**% on **domain-specific extraction** (e.g., dosing, covariates) on pharma domain datasets and **10**% on other more **general** type extractions evaluated on 150 various sources. It is also **10**% more accurate when generating large tables with $>$ **100** rows. [Demo]
  - The pipeline creates several fields to extract from user queries and provides the data through a structured table format and ensures correct data type.
- **Engineered** an advanced paper layout system to **enhance Retrieval-Augmented Generation (RAG)** performance by **12**%. Key system capabilities include:
  - Trained a sub-figure model (**Global Average IoU: 0.8919**) based on YOLOv11 to decompose composite figures into **subfigures** and extract metadata (captions, legends), significantly boosting **visual IR**.
  - Augmented all image, and table data chunks with **section-level location metadata**, captions, footnote and summaries to which enabled sophisticated search via **image queries** (e.g., "Find breast cancer papers that contain drug name **on the x-axis** and dosage amount **on the y-axis**").
- Created a production-grade qdrant vector database containing **40M** paper pdfs for mass scale paper search.
- **Engineered** a **novel image Bounding Box (BBox) filtering algorithm** that significantly enhances data quality by reducing false positives by **34**%.
  - Integrated **dynamic minimum bbox size** based on page heuristics into NMS filtering to automatically eliminate erroneous small BBoxes.
- Collaborated with pharmaceutical scientists to create an **automated QC** system for cleaning and fixing pharmaceutical data (unit standardisation, column standardisation, biomarker checking, dosing range validation, etc.), improving data accuracy to **97**%.
- **Established Scalable MLOps and Secure Deployment Infrastructure** for high-throughput inference services.
  - Optimized custom ML model inference latency and throughput using **Nvidia TensorRT** compilation and deployment via the **Nvidia Triton Inference Server (TIS)** framework.
  - Migrated existing long-running tasks to **Celery**, implementing an event-driven architecture using **RabbitMQ (RMQ)** as the event bus for decoupled data transfer.
  - **Engineered event-driven autoscaling in Kubernetes (K8s)** using **KEDA** to monitor **RMQ queue length**, dynamically scaling Celery worker pods to maintain a consistent tasks-per-worker ratio and ensure near-zero lag.
  - Enforced a **Zero Trust Security** architecture by migrating all public cloud services to private networks, restricting service-to-service communication via **VPC Endpoints** and **secure S2S tokens**.

Jr AI Engineer — Sep. 2022 – Sep. 2024
Giga Tech Ltd. — *Gulshan, Dhaka, Bangladesh*

- **Achieved SOTA in Low-Resource Bangla NLP (NER, POS, QA, Lemmatization)**, exceeding **90**% KPI for all modules.
  - Set new **NER SOTA** (81.85% Macro F1, **+6%** vs. prior SOTA, 90.49% Macro F1 on delivererd dataset,) & **POS SOTA** using a novel hierarchical voting mechanism among **SLM** predictions. [Demo]
  - Delivered **SOTA for QA** (SQuAD-bn) with a novel loss function to balance null/non-null answers, beating prior SOTA by **6**%. [Demo]
  - Co-developed & open-sourced the first production-grade Bangla **Lemmatizer** (96.36% accuracy) and **Emotion Recognition** system. [Github, Demo].
- **Engineered Advanced LLM Methodologies and Data Pipelines** for complex generation and tagging tasks.
  - Built **GPT-4o inference pipelines** (NER, Coref) using the **ReAct** prompting framework, matching fine-tuned model performance.
  - Resolved severe class imbalance in sequence tagging via a **sentence resampling** pipeline and advanced loss functions (**Dice, Focal, CurricularFace**).
- **Architected Core MLOps and Data Versioning Infrastructure** to manage the complete model/data lifecycle, leveraging **DVC**, & **MLflow**.

PUBLICATIONS

BanNERD: Context-Driven Approach for Bangla Named Entity Recognition 2024
2025 Conference of the Nations of the Americas Chapter of ACL (NAACL); H-Index:218 *New Mexico*
- Developed **BanNERCEM**, a novel context-ensemble method achieving a **state-of-the-art (SOTA) macro F1 score of 81.85%** on Bangla NER, outperforming previous approaches.

BanLemma: A Word Formation Dependent Rule Based Lemmatizer 2023
The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP); H-Index:176 *Singapore*
- Developed **BanLemma**, a novel rule-based lemmatizer achieving **96.36% accuracy** on a manually annotated dataset by analyzing suffix markers from a 90M sentence corpus. [Source: Paper, Code]

BanglaBait: Semi-Supervised Adversarial Approach for Clickbait Detection 2023
International Conference on Recent Advances in Natural Language Processing; H-Index:36 *Varna, Bulgaria*
- Created the **first Bangla Clickbait Dataset (15k instances)** and demonstrated the superiority of semi-supervised learning methods for this task. [Source: Paper, Code]

Technical Skills

Programming Languages: Python, JavaScript, Bash, SQL
ML/DL Frameworks & Libraries: PyTorch, Hugging Face Transformers, LangChain, LangGraph, LlamaIndex, OpenCV,
MLOps & Experimentation: MLflow, DVC, Weights & Biases, Langfuse, Prometheus, Grafana, Ray, Dask, Celery, RabbitMQ
Model Optimization & Deployment: Triton, ONNX
Cloud & DevOps/IaC: AWS, AWS Sagemaker, GCP, Docker, Kubernetes (K8s), Terraform, Git
Databases & Vector Stores: PostgreSQL, Elasticsearch, Neo4j, Qdrant, FAISS
Web Development & APIs: FastAPI, Flask, Streamlit
Data Engineering & ETL: Apache Kafka, Apache Spark, Apache Airflow
Testing & Quality Assurance: Locust (Load Testing)
Specialized Tools (LLMs/RAG): N8N, Codex

Academic Achievements

BRAC University Intra University Programming Contest | *Winner* 2019
BRAC University Dean's Prestigious List Award 2022

Articles

- **Medium** — Sparse Transformers Explained — URL: medium.com/@mahtab27672767/sparse-transformers-explained-part-1-aacbe10dca4a

Open Source Contributions

- https://github.com/flairNLP/flair/pull/3449