Aritra Kumar Lahiri

Ph: +1 5142102763 | E-mail: aritra.lahiri@torontomu.ca | LinkedIn | GitHub | Portfolio

Summary: Seasoned AI/ML Engineer with 8+ years of industry experience in software engineering and applied research. With extensive research background in LLMs, Multimodal RAG, Information Retrieval, and NLP, for building real-world applications, I have actively contributed to emerging AI trends, agentic AI workflows and high-impact AI innovations. I have mentored junior researchers and led research in multimodal RAG apps in clinical and narrative domains with publications in top-tier journals/conferences such as IEEE, ACM, ECIR, CIKM, Springer.

TECHNICAL SKILLS

- **Programming Languages & Databases:** Java, Python, SQL, JavaScript, Hive, NodeJS, MongoDB, Postgres
- **Methodologies**: Spring MVC, Spring Boot, RESTful APIs, Microservices, Kafka, RAG, OAuth2, JSON
- **Machine Learning:** PyTorch, Keras, Scikit-Learn, TensorFlow, HuggingFace, Transformers, Flask, FastAPI, Jinja2 Templates, LLMs, PEFT, QLoRA, Model Benchmarking, Question-Answering, Sentiment Analysis
- **Foundation Tools:** HTML, CSS, Angular, Gradle, Maven, LangChain, FaissDB, NumPy, Pandas, Matplotlib, NLTK, Gensim, Jupyter Notebook, Git
- **Infrastructure**: Docker, GCP Apigee, MLflow, Jenkins CI, SonarQube, Agile-Jira, Confluence

PROFESSIONAL EXPERIENCE

AI Engineer, Mercor, Remote (October 2025 - Present**)**

- Collaborated with AI research lab to develop evaluation framework for advancing State-of-the-Art model understanding and performance.
- Optimized and refined client-specific prompts for agentic framework establishing best practices.
- Integrated LLMs into Python-based Automated evaluation framework to validate conversational AI workflows.

Applied AI Engineer, Vector Institute, Toronto, ON, CA (May 2025 – September 2025**)**

- Led development of a multimodal RAG pipeline combining text and image embeddings from PubMed clinical documents using LLaVA and LLaMA.
- Developed semantics-aware document chunking pipeline to optimize data processing for long-form documents.
- Integrated visual spatial cues (e.g., image-derived context vectors) with textual retrieval to support clinical decision-making with multimodal foundation model use cases.
- Improved hallucination control and spatial alignment using cross-attention fusion and ROUGE/BLEU metrics.

Software Engineer II, TD Bank, Toronto, ON, CA (June 2022 – May 2024**)**

- Led design and development for payment and digital credit offer API using SpringBoot, Java and IBM DB2 database server for enhancing TD Easy Web offer acceptance workflow.
- Spearheaded junior developers to establish the API framework standards template automation and custom logger flow in Java and NodeJS.
- Achieved success in implementing the end-to-end API developer portal workflow customizing GCP Apigee policies and deploying through MS Azure.

Software Engineer, Ford Motors, Detroit, MI, USA (June 2019 – May 2022**)**

- Implemented Rest Api using Spring Data JPA in Java to query Hive tables.
- Integrated IIoT platform with secure ML APIs for anomaly detection and data transmission encrypted with internal token guards, saving huge business cost, won Tech award for innovation.
- Integrated Python ML model with Alteryx workflow utilizing Rest API and optimized batch scheduler in Kafka message queue for callback response microservice and audit logging.
- Tools - Angular 6, SpringBoot in Java at backend, MS SQL server, PCF, AWS S3, LDAP

Software Development Engineer, Pearson, Chandler, AZ, USA (June 2016 – May 2019**)**

- Developed REST APIs utilizing Spring and Hibernate for persistence and fixed integration endpoint mapping following open API spec in Swagger.
- Designed and implemented Spring Boot Microservices for search and indexing data into backend MongoDB.

PROJECTS

AlzheimerRAG: Multimodal Retrieval Augmented Generation for clinical use cases using PubMed Articles

- Multimodal RAG application for biomedical/clinical research use cases (Alzheimer's) from PubMed articles.
- Incorporated multimodal fusion techniques to integrate unstructured data through textual and visual data processing by applying embedding vector integrity checks using FaissDB and GPT-based validation for retrieved evidence.
- Improved retrieval precision by 25% and LLM response coherence by 20% by upgrading from traditional to multimodal RAG.
- Achieved 10x faster inference using QLoRA model quantization and custom re-ranking model while maintaining high F1-score (86%) on PubMedQA over existing benchmarks.
- Tools & Technology: LLMs such as LlaMA and LlaVA, LangChain, FaissDB, Jinja2, FastAPI, GPT-4.0.

DragonVerseQA: Long-Form Context-Aware QA with Knowledge Graph Alignment

- Curated multidimensional narrative domain-specific dataset of 3000+ Open-Domain Long-Form Context-Aware QA by integrating text summaries, user sentiment, and structured knowledge graphs.
- Constructed long-form answers integrating Wiki Data, IMDb, and episodic user behaviour.
- Incorporated semi-supervised learning for spam and bias filtering, entity salience check, and data pruning for QA pair robustness.
- Tools & Technology: LLMs like fine-tuned variants of GPT, SVM, Neo4J for knowledge graphs, BERT, Random Forest, SVG, Zero-shot and Few-shot Prompting, Web-Scraper, Python, NLP Evaluation libraries.

TREC Clinical Trials Data Retrieval

- Performed data extraction using PubMed Parser combining cleaning and data preprocessing for model inputs.
- Retrieved clinical trial entries from PubMed using Doc2Vec and caption-augmented rankers.
- Evaluated relevance using NDCG and cosine similarity metrics.
- Tools & Technology: Python, Sentence Transformer, Doc2Vec, NLTK, TF-IDF vectorizer, NDCG

Auto-Answer Aware QA generator

- Developed Answer-Aware QA generator using deep learning and secure API serving by leveraging transfer learning techniques through neural language models to simplify the question-answer generation.
- Curated a corpus of 10,000+ QA pairs on Game of Thrones Series using Answer-Aware Question Generation
- Implemented as a Python and Flask-based web application that provides user interface to visualize the QA pair generation task and serves the client request by rendering the final output via API calls.

Transformer-Based Text Summarization Microservice

- Fine-tuned Google Pegasus transformer model on extracted data from “A Song of Ice and Fire” book using HuggingFace Trainer, achieving significant improvement in ROUGE-L and enhancing short-form text abstraction accuracy in domain-specific task.
- Built a full-stack MLOps pipeline encompassing data ingestion, model training, evaluation, and CI/CD integration with Docker and GitHub Actions, reducing deployment time by 50%.
- Deployed the summarization model as RESTful microservice on GCP cluster with FastAPI and Jinja2 templates, enabling real-time user interaction and feedback-driven model refinement.

Inspection Ratings of Chicago Restaurants

- ETL workflow for calculating food inspection ratings on Chicago Restaurants.
- Tools & Technology: Pandas for data munging, SQL for storing transformed data and Python, Mongo and Flask App for the load, Kaggle Dataset source.

ACADEMIC QUALIFICATION

PhD in Computer Science Toronto Metropolitan University, Toronto, ON, CA, Completion – October 2025
MS in Computer Science, Arizona State University, Tempe, AZ, USA, Year of Completion - 2016

PUBLICATIONS

- A. K. Lahiri and Q. V. Hu, “Descriptor: Open-Domain Long-Form Context-Aware Question-Answering Dataset (DragonVerseQA)," in IEEE Data Descriptions, doi: 10.1109/IEEEDATA.2025.3562173.
- Lahiri, Aritra Kumar, and Qinmin Vivian Hu. 2025. "AlzheimerRAG: Multimodal Retrieval-Augmented Generation for Clinical Use Cases" *Machine Learning and Knowledge Extraction* 7, no. 3: 89. https://doi.org/10.3390/make7030089.
- Lahiri, A. K., & Hu, Q. V. HouseOfTheDragonQA: Open-Domain Long-Form Context-Aware QA Pairs for TV Series. In 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp 150 - 157)
- Lahiri, Aritra Kumar, and Qinmin Vivian Hu. "GameOfThronesQA: Answer-aware question-answer pairs for tv series." European Conference on Information Retrieval (ECIR). Cham: Springer International Publishing, 2022.
- Lahiri, A. K., Hasan, E., Hu, Q. V., & Ding, C. (2023). TMU at TREC Clinical Trials Track 2023. In I. Soboroff & A. Ellis (Eds.), The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14–17, 2023 (NIST Special Publication No. 500-xxx). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec32/papers/V-TorontoMU.C.pdf
- Lahiri, A. K., & Hu, Q. V. Entity Level QA Pairs Dataset for Sentiment Analysis. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (pp. 270-276).
- Lahiri, Aritra Kumar, and Q.V. Hu. "Named Entity-based Question-Answering Pair Generator." In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 4902-4906. 2022.
- A Game Theoretic Approach to Demand Side Management in Smart Grid with Multiple Energy Sources and Storage. *International Journal of Advanced Computer Science and Applications (IJACSA), The Science and Information (SAI) Organization* Volume 9, Issue 2, February 2018