Aritra Kumar Lahiri

Toronto, ON, CA
Ph: +1 5142102763; E-mail: aritra.lahiri@torontomu.ca
LinkedIn: https://www.linkedin.com/in/aritrakumarlahiri/ GitHub: https://github.com/Aritra23
Portfolio: https://aritra23.github.io/

Summary: PhD Candidate and AI Researcher with 9+ years of software and applied research experience in context-aware QA systems, multimodal foundation models specializing in LLM/VLM fusion, RAG pipelines, and deep learning for real-world applications. Published in top-tier conferences (IEEE, ACM, ECIR, Springer), I have led research in multimodal QA and visual grounding in clinical and narrative domains. Skilled in PyTorch, TensorFlow, LLaVA, and GPT-family models, with a strong record of integrating research prototypes into production environments using GCP, Docker, and FastAPI. Actively mentored junior researchers and contribute to high-impact, IP-focused AI innovations.

TECHNICAL SKILLS

- **Languages:** Python, Java, SQL, JavaScript, MongoDB, NodeJS
- **Frameworks**: Spring MVC, Spring Boot, RESTful APIs, Microservices, AWS, NodeJS, React, GCP, Apigee
- **Deep Learning:** PyTorch, Keras, Scikit-Learn, TensorFlow, HuggingFace, Transformers, Flask, FastAPI, Jinja2 Templates, LLMs
- **Foundation Tools:** LangChain, FaissDB, NumPy, Pandas, Matplotlib, NLTK, Gensim, Spacy
- **Infrastructure**: GCP, Docker, Apigee, FastAPI, Jupyter Notebook, Git

PROFESSIONAL EXPERIENCE

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

- Implemented payment and digital credit offer API using SpringBoot, Java and IBM DB2 database server for enhancing TD Easy Web offer acceptance workflow, one of its primary user enablement products.
- Spearheaded a group of developers to establish the API framework standards in Java and NodeJS.
- Achieved success in developing the end-to-end API developer portal solution using GCP Apigee policies. Utilized Python scripts for designing the CI/CD pipelines

Software Engineer, Ford Motors, Detroit, MI (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.
- Developed batch pipelines with callback response isolation and audit logging for real-time manufacturing data.
- Tools - Angular 6, SpringBoot in Java at backend, MS SQL server, PCF, AWS S3, LDAP and Email as micro-services.

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

- Developed REST APIs utilizing Spring and Hibernate for persistence. Fixed integration endpoint mapping following open API spec in Swagger, consumed REST APIs using Swagger UI.
- Designed and implemented Spring Boot Microservices for search and indexing data into backend MongoDB.
- Used JUnit for TDD integrated with Ant, Jenkins, improved unit test coverage by 80 %, collaborated through Bitbucket.

INTERNSHIP

Applied AI Intern, Vector Institute, Toronto (May 2025 – Present**)**

- Led development of a multimodal RAG pipeline combining text and image embeddings from PubMed clinical documents using LLaVA and LLaMA.
- 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.
- Collaborated across teams to build scalable APIs (FastAPI, LangChain) and optimize pipeline latency.

Web Developer Intern, Arizona State University. Scottsdale, AZ (May 2015 – August 2015**)**

- Prototyped a health analytics recommender system, that maps disease risk factors with biosensor device features and generate a mapping algorithm to prescribe devices based on patient profile.
- Designed MVC architecture using Angular JS at front end, Node JS Server, Express JS for connecting with the DB at back end and MongoDB (NoSQL) database. Installed Ruby for running Sass CSS, Compass, Git for version control.

ACADEMIC & RESEARCH PROJECTS

Title: AlzheimerRAG: Multimodal Retrieval Augmented Generation for PubMed Articles
Publication: July 2025
Project Summary: Multimodal RAG application for biomedical research use cases, primarily focusing on Alzheimer's disease from PubMed articles. The application incorporates multimodal fusion techniques to integrate textual and visual data processing by applying for embedding vector integrity checks using FaissDB and GPT-based validation for retrieved evidence.

Tools & Technology: LLMs such as LlaMA and LlaVA, LangChain, FaissDB, Jinja2, FastAPI, GPT-4.0.

Title: DragonVerseQA: Context-Aware QA with Knowledge Graph Alignment
Publication: January 2025
Project Summary: Open-Domain Long-Form Context-Aware QA dataset based on fantasy universe genre of TV series “House of Dragons” and “Game of Thrones”. Developed context-aware QA over multimodal narrative domains (GoT, House of Dragons), integrating text summaries, user sentiment, and structured graphs. Constructed long-form answers integrating WikiData, IMDb, and episodic user behaviour. Incorporated bias filtering, entity salience check, and data pruning for QA pair robustness.
Tools & Technology: LLMs like fine-tuned variants of GPT, SVM, Neo4J for graphs, Web-Scraper, Python, NLP Evaluation libraries.

Title: TREC Clinical Trials Track
Publication: November 2023
Project Summary: Retrieved visual clinical trial entries from PubMed using Doc2Vec and caption-augmented rankers. Evaluated relevance using NDCG and cross-modal similarity metrics.
Tools & Technology: Python, SentenceTransformer, Doc2Vec, NLTK, TF-IDF vectorizer, NDCG

Title: Auto-Answer Aware QA generator
Publication: October 2022
Project Summary: Developed QA generator using transfer learning and secure API serving by leveraging transfer learning techniques to simplify the question-answer generation. The tool utilizes the Answer Aware Question Generation (AAQG) method using the fine-tuned transformer-based model and task-based prefixes to generate high-quality QA pairs. 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.
Tools & Technology: AAQG model (fine-tuned BERT-HLSQG), Python, Flask.

Title: GameOfThronesQA: Answer-Aware Question-Answer Pairs for TV Series
Publication: April 2022
Project Summary: Curated a corpus of QA dataset using Answer Aware Question Answering techniques and provided a novel pipeline framework for Answer Aware question Generation.
Tools & Technology: LLMs such as fine-tuned variant of BERT, Python.

Title: Implementation of Data Partitioning techniques, Query processing operations
Duration: August 2015 – December 2015
Project Summary: Analysed Movie-Lens Dataset with 10M records to implement range partitioning and round – robin partitioning. Implemented parallel sort and parallel join operations on partitioned tables.
Tools & Technology: Python and PostgreSQL

INDEPENDENT PROJECTS

Title: Spotify Data Visualization
Duration: March 2019
Team Size: 3
Project Summary: Analyzed Spotify Million Playlist Dataset to extract and visualize top artists and tracks by each region.
Tools & Technology: JVectorMaps(native JS library), Python, Pandas, Jupyter Notebook, Plotly.JS.

Title: Mission to Mars
Duration: January 2019
Team Size: 1
Project Summary: Implemented Flask App to render scraped Mars related data from NASA website and display the information in a single page application.
Tools & Technology: Jupyter Notebook, BeautifulSoup, Pandas, Splinter, PyMongo and Flask.

Title: DOM Manipulation using UFO Sighting Dataset
Duration: January 2019.
Team Size: 1
Project Summary: Manipulate Document Object Model to filter dataset based on single or multiple search categories. **Tools & Technology: Python**, JS.

Title: Data Journalism using D3
Duration: January 2019
Team Size: 1
Project Summary: Interactive Data Visualization using D3.JS on U.S. Census Bureau and the Behavioral Risk Factor Surveillance System dataset.
Tools & Technology: D3.JS

Title: Inspection Ratings of Chicago Restaurants
Duration: January 2019
Team Size: 3
Project Summary: 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.

Title: Climate Analysis and Exploration
Duration: December 2018
Team Size: 1
Project Summary: Analysis of Weather across 500+ cities comprising of different parameters like cloudiness, humidity, max temperature, and wind speed.
Tools & Technology: Python, SQL Alchemy ORM queries, Pandas, Matplotlib and Open Weather Map API.

Title: Mini Messaging Service (Twitter)
Duration: December 2017
Team Size: 1
Project Summary: Implemented a backend RESTful API of a mini message service like Twitter supporting features like view, write, search tweets, follow and unfollow users as well as list of all users following the current user and their most popular follower.
Tools & Technology: Spring Boot, IntelliJ IDE, Java, H2 database

ACADEMIC QUALIFICATION

PhD in Computer Science Toronto Metropolitan University, Toronto, ON, CA, Completion – August 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, 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

TRAINING & CERTIFICATIONS

- Data Analytics from University of Arizona, Tucson, Phoenix, 2018.
- Machine Learning from Coursera 2018