SAYAK BANERJEE

Email: sayakbanerjee022@gmail.com

Phone Number: (412) 844-1988

LinkedIn/sayak-banerjee/

Google-Scholar/sayak-banerjee/

EDUCATION

Carnegie Mellon University   **Pittsburgh, PA**
Master of Computational Data Science (MCDS) | QPA – 4.00/4.00   **Dec 2025**
Coursework – Large Language Models, Search Engines, Introduction to Machine Learning, Cloud Computing
Teaching Assistant – Fundamentals of Operationalizing AI, Python for Data Science

Vellore Institute of Technology   **Vellore, India**
B. Tech, Electronics and Communication Engineering | CGPA – 9.41/10   **May 2022**
Relevant Coursework – Data structures, Algorithms, Operating Systems, Database Systems, Statistics, Applied Linear Algebra

SKILLS

Languages & Tools: Python, C++, PySpark, MySQL, Snowflake, Git, Apache Airflow, Docker, Kubernetes, Helm, AWS, Azure
Libraries and Frameworks: PyTorch, Huggingface, Tensorflow, NumPy, Pandas, SciPy, Scikit Learn, Matplotlib, SpaCy
Technologies: Machine Learning, Natural Language Processing, Large Language Models (**LLM**) & Foundation Models, Information Retrieval, RAG, Deep Learning Systems, Cloud Computing, ETL Pipelines, CI/CD, Containerization

EXPERIENCE

Paylocity   **Schaumburg, IL**
Data Science Intern   **May 2025 – Present**
- Collaborating with the Search team to enhance an enterprise chatbot solution utilizing RAG architecture focussing on IRS-compliance and company handbook related queries.
- Improved the existing re-ranking pipeline by removing BM25, which resulted in an average **reduction of 5k tokens** across user queries and **reduced latency by ~2 seconds**. Further, enhanced the initial ranking pipeline by incorporating and finetuning multi-match keyword-based search alongside **HNSW ANN search**, creating a hybrid ranking pipeline.

Acuity Knowledge Partners   **Gurugram, India**
Associate (Data Engineer)   **Dec 2022 – July 2024**
Senior Analyst   **July 2022 – Nov 2022**
- Orchestrated and analyzed **alternative financial data** for a US-based Hedge Fund covering various investment portfolios including consumer, healthcare, technology, and e-commerce.
- Spearheaded the integration of Apache Airflow with existing ETL pipelines – creating event-based and time-based triggers, reducing manual efforts to monitor jobs **by more than 60%**.
- Designed ETL pipelines and implemented a dynamic data warehouse utilizing Python, PySpark, Pandas, Airflow, AWS, SQL, and Snowflake to store aggregated KPI data from multiple datasets in a SQL database, empowering PMs **to generate financial models by automating the extraction of over 1 million data points**.
- Implemented a full-fledged analytics pipeline involving product tagging and KPI visualizations using PySpark, Pandas, and Dash, **achieving ~40% reduction in time taken to generate KPI reports.**

Bajaj Finance Limited   **Pune, India**
Data Science and Engineering Intern   **Jan 2022 – June 2022**
- Developed various customer segmentation and propensity models using machine-learning algorithms for the health insurance business. **Accomplished a ~6% overall increase in effective lead generation of prospective customers**.

PROJECTS

Carnegie Mellon University   **Pittsburgh, PA**
Optimizing QA performance with Dense Passage Retrieval (DPR) and RAG   **April 2025**
- Developed a modular multi-stage response generation pipeline (DPR → Reranking → RAG) integrating dense passage retrieval on FAISS-indexes and Flan-T5 based QA agent, enabling end-to-end neural retrieval and answer generation.
- Integrated Lucene for efficient lexical indexing and FAISS with co-condenser architecture for DPR.
- Conducted extensive experiments combining advanced prompting techniques (CoT, Persona prompting), Learning to Rank features with diverse reranking pipelines (BM25, LTR-based SVMRank/Coordinate Ascent/ListNet, BERT-*n*), **achieving a 25% improvement in exact match (EM) scores** over the BM25 baseline on the SQuAD evaluation dataset.

Twitter Recommendation Microservice (A Cloud-Native Scalable Solution)   **April, 2025**
- Engineered a distributed ETL pipeline using PySpark on Databricks to process ~1TB of raw Twitter JSON data; computing scores based on user interactions, keywords, and hashtags for user-recommendation.
- Deployed a fault-tolerant production-grade microservice with MySQL back-end (optimized using denormalization and indexing) on a self-managed Kubernetes cluster (AWS ECS) using docker, helm charts; implemented automated CI/CD pipelines with GitHub actions. The microservice achieved 20K+ RPS over a 3-hour live load test.

Text2Utility – Talk with your Mobile Apps   **Dec, 2024**
- Developed a LLM-based AI agent for end-to-end task automation from natural language inputs.
- It involves intent classification, structured code generation (JSON) followed by LLM tool-use to help user's **complete tasks such as booking an Uber, finding a restaurant table or getting weather updates**.
- Conducted extensive model evaluation, implementing PEFT (LoRA) along with few-shot learning to benchmark various foundation models from Hugging Face, quantifying task-specific accuracy metrics across different user intents.