SAYAK BANERJEE

Email: sayakbanerjee022@gmail.com LinkedIn/sayak-banerjee/
Phone Number: (412) 844-1988 Google-Scholar/sayak-banerjee/

EDUCATION

Carnegie Mellon University **Pittsburgh, PA**
Master of Computational Data Science (MCDS) | QPA – 4.00/4.00 **Dec 2025**
Coursework – 11-667: Large Language Models, 11-642: Search Engines, 10-601: Machine Learning, 15-619 Cloud Computing

Vellore Institute of Technology **Vellore, India**
B. Tech, Electronics and Communication Engineering | CGPA – 9.41/10 **May 2022**

SKILLS

Languages & Tools: Python, C++, PySpark, SQL, Snowflake, Git, Apache Airflow, MLFlow, Kafka, Docker, Kubernetes, AWS
Libraries and Frameworks: PyTorch, Huggingface, Tensorflow, NumPy, Pandas, SciPy, Scikit Learn, Matplotlib, SpaCy
Technologies: Machine Learning, ML Algorithms, Natural Language Processing, Large Language Models & Foundation Models, Search & Information Retrieval, RAG, Deep Learning Systems, Time Series, Cloud Computing, ETL Pipelines, Vector Databases

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 queries related to the Paylocity portal, IRS-compliance and company handbooks.
- Implemented a dynamic fallback mechanism to switch to Bedrock embeddings during OpenAI latency spikes or outages, after thorough experimentation and benchmarking – **improving the AI assistant's SLA by ~2% for our customers**. **This work laid the foundation for upcoming load balancing and model routing enhancements.**
- 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 ~2.5 seconds**.
- Enhanced the ranking pipeline by finetuning multi-match keyword-based search alongside **HNSW ANN search**, creating a hybrid ranking pipeline, **improving recall by 0.6% and acceptable answer proportion by 1.44%**.

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**.
- Developed Time Series forecasting models to estimate the growth of data internally for storage requirements – allowing to switch to dynamic pay-as-you-go models on AWS – **reducing storage and volume costs by nearly 15%.**
- 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.**

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.