SAYAK BANERJEE

+1 (412) 844-1988 | sayakbanerjee022@gmail.com
LinkedIn/sayak-banerjee99/ | Google-Scholar/sayak-banerjee/ | Github/sayakbanerjee1999

EDUCATION

Carnegie Mellon University **Pittsburgh, PA**
Master of Computational Data Science (MCDS), ***QPA – 4.05/4.00*** **Dec 2025**
Courses – Machine Learning, Large Lang. Models, LLM Inference, Search Engines, Deep Learning Systems, 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, Docker, Kubernetes, Git, Airflow, MLFlow
ML & AI Frameworks: PyTorch, TensorFlow, Hugging Face, Lang Chain/Graph, vLLM, Scikit-Learn, Pandas
ML Specialization: Machine Learning, Natural Language Processing, Large Language Models, Multimodal Models, AI Agents, Retrieval-Augmented Generation (RAG), Information Retrieval, Search Engines, Vector Databases, Deep-Learning Systems
Data and Systems: AWS (EKS, ECS, Bedrock, RDS), Inference Engineering, Model Evaluation, Apache Kafka, CI/CD, ETL

EXPERIENCE

Acuity Analytics **New York City, NY**
Quantitative Research Associate **Jan 2026 – Present**

- L/S Equity Hedge Fund
- Implementing data science specific back-testing frameworks for alternative financial datasets – driving data-driven validation of alpha signals and supporting firm-wide AI initiatives and quantitative research.

Paylocity **Schaumburg, IL**
Machine Learning Engineer Intern **May 2025 – Aug 2025**

- Optimized a RAG-based enterprise AI assistant through advanced ANN search techniques, improving query relevance and performance for questions related to the Paylocity portal, IRS compliance and company handbooks.
- Redesigned the ranking pipeline by replacing BM25 with a **hybrid ranking pipeline (multi-match keyword search + HNSW ANN search)**, **reducing average query cost by ~5k tokens** and **latency by 2.5s**, while **improving recall by 0.6%** and **acceptable answer proportion by 1.44%**.
- Implemented a dynamic fallback mechanism to switch to Bedrock embeddings during OpenAI latency spikes or outages – **improving SLA by ~2%.** This work laid the foundation for load balancing and model routing enhancements.
- Architected a workflow to detect and extract tables from table heavy pdfs using table transformers and GPT-4o mini and stored them in vector databases in an uniform markdown format for enhanced retrieval and query performance.

Acuity Knowledge Partners **Gurugram, India**
Associate - Data Science and Engineering **July 2022 – July 2024**

- 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 on structured and un-structured datasets, **reducing manual efforts to monitor jobs by more than 60%.**
- Designed ETL pipelines and implemented a dynamic data warehouse using Python, PySpark, Pandas, Airflow, AWS, SQL, and Snowflake to store aggregated KPI data from multiple vendor datasets in a SQL database, empowering PMs **to generate financial models by automating the extraction of over 1 million data points**.

PROJECTS

Carnegie Mellon University **Pittsburgh, PA**
Advanced LLM Inference Server across Shared Tasks **Nov 2025**

- Implemented a production-grade inference server and benchmarked advanced decoding algorithms for LLM inference, including **Speculative decoding**, **K-V caching**, **Mirostat, diverse beam search, continuous batching with disaggregated prefill and decode** – optimizing generation efficiency for Qwen models across distributed workloads.
- Deployed a **Self-Refine** inference pipeline, implementing iterative generation-critique-refinement loops with **tool calling integration** for GraphDev, achieving measurable gains in accuracy across GraphDev and MMLU-Med datasets.

Optimizing QA performance with Dense Passage Retrieval (DPR) and RAG **April 2025**

- Developed a **modular multi-stage QA 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.

Text2Utility – Talk with your Mobile Apps **Dec 2024**

- Developed an AI agent for task automation from natural language inputs, combining intent classification, structured code generation and LLM tool-use to execute personalized tasks such as **booking an Uber, finding restaurant tables**.
- Benchmarked Hugging Face foundation models with **Parameter Efficient Fine Tuning (LoRA)** and few-shot learning, **achieving 12% gains in task-specific accuracy** across diverse user intents compared to the baseline BERT-large model.