

SAYAK BANERJEE

Pittsburgh, PA | (412) 844-1988 | sayakbanerjee022@gmail.com | [LinkedIn/sayak-banerjee/](#) | [Google-Scholar/sayak-banerjee/](#) | [Github/sayak-banerjee](#)

EDUCATION

Carnegie Mellon University

Master of Computational Data Science (MCDS), QPA – 4.00/4.00

Pittsburgh, PA

Dec 2025

Coursework – Machine Learning, Large Language Models, LLM Inference, Search Engines, Deep Reinforcement Learning and Control, Deep Learning Systems, Cloud Computing

Vellore Institute of Technology

B. Tech, Electronics and Communication Engineering, CGPA – 9.41/10

Vellore, India

May 2022

SKILLS

Languages & Tools: Python, C++, PySpark, SQL, FAISS, Snowflake, Docker, Kubernetes, Git, Airflow, MLFlow, Kafka, AWS

ML & AI Frameworks: PyTorch, TensorFlow, Hugging Face, LangChain/Graph, vLLM, Scikit-Learn, XGBoost, Pandas

ML Specialization: Machine Learning, Natural Language Processing (NLP), Large Language Models (LLM), AI Agents, Retrieval-Augmented Generation (RAG), Information Retrieval, Search Engines, Vector Databases, Deep-Learning Systems, Multimodal Models, Convolutional Neural Networks (CNNs), Feature Engineering

Data and Systems: Production ML, Optimization, Model Evaluation, Cloud Computing, A/B Testing, CI/CD, ETL Pipelines

EXPERIENCE

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 the AI assistant's SLA by ~2%**. This work laid the foundation for the upcoming load balancing and embedding 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**.
- 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

Advanced LLM Inference Algorithms across Shared Tasks

Nov 2025

- Implemented and benchmarked advanced decoding algorithms for LLM inference, including **Speculative decoding, KV caching, Mirostat, diverse beam search** – optimizing generation efficiency for Qwen model variants across Infobench, MMLU-med and GraphDev datasets.
- Built 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.