Santanu Banerjee *Data Scientist at Capgemini*

iamsantanubanerjee.github.io iamsantanubanerjee in/iamsantanubanerjee contactsantanubanerjee@gmail.com

+91 90511 41779 Kolkata, India

SUMMARY

Data scientist with 2+ years of experience in applying advanced machine learning, deep learning and generative ai models to extract insights from complex datasets. Possesses a strong foundation in mathematics and a demonstrated aptitude for cutting-edge technologies through internships focused on quantum computing and quantum machine learning.

PROFESSIONAL EXPERIENCE

Capgemini, *Data Scientist* Dec 2021 – present | Kolkata, India

- **Intelligent Knowledge Management and Retrieval System** (A RAG-based GenAI Q&A platform for enterprise knowledge base)
  - **Data Processing Pipeline:** Developed a robust pipeline using Azure Document Intelligence and Azure Machine Learning Studio to extract data from various document formats (PDFs, presentations). Implemented page-wise extraction and chunking algorithms to efficiently store data in JSON format.
  - **Metadata Generation and Embedding:** Utilized Azure OpenAI's text-embedding-3-large model to create metadata and embeddings, enhancing search capabilities. Metadata included document ID, content, file name, file path, page number, and embedding vectors.
  - **Semantic Search and Indexing:** Employed Azure Search and Azure Cognitive Search to index and ingest metadata into a vector database, enabling efficient semantic search via cosine similarity.
  - **Evaluation and Citation:** Orchestrated the evaluation process by querying the vector database, comparing results with ground truth questions, and tracing relevant sources using Azure Search, Azure OpenAI, and Azure Machine Learning Workspace.
  - **Integration with Language Models and Evaluation of RAG:** Integrated Azure Open AI GPT 3.5 turbo 16k with the pipeline to generate model responses based on user queries and relevant contexts along with prompts. Developed evaluation metrics utilizing RAGAS architecture to assess system performance in terms of faithfulness, answer relevancy, context relevancy and answer correctness based on user query, ground truth, model response and relevant context with different chunk size, dimensions and chunk overlap.
  - **End-to-End Application Development:** Created a modular, scalable application using Streamlit and Azure VS Code, enabling seamless interaction with the document processing and retrieval system.
- **SecureNet AI Watchtower** (A RAG-based GenAI platform for your enterprise to attack potential privacy and security concerns)
  - **Vectorized Data Storage and Retrieval**: Leveraged LlamaIndex to convert large amount of sensitive data (employee chat conversations in the form of emails, texts, audio, video, pdfs), which were previously involved with data leakage, into vectorized representations and stored them efficiently in MongoDB Atlas vector database.
  - **Risk Assessment through Cosine Similarity**: Implemented a cosine similarity-based approach to calculate a risk score by comparing new conversations with the most relevant vectorized conversations (retrieved from the vector database based on keywords provided by the user). This risk score served as an indicator for potential data leakage in these new conversations.
  - **Automated Monitoring and Actionable Insights**: Integrated open-source Large Language Model (Llama2) for generating summaries from conversations that exceeded a certain risk score threshold and extracting relevant participant information from those conversations. The processed data was then visualized on a PowerBI dashboard, empowering the monitoring team with actionable insights to identify and address potential risks promptly.
- **Data Archival to OpenText InfoArchive**
  - Developed ETL-based data archiving solution using Python scripting to facilitate the implementation of data archival processes, ensuring long-term preservation and access to critical data.
  - Collaborated with cross-functional teams to identify and prioritize data to be archived, determine appropriate storage locations and formats, maintained documentation which included policies, procedures and workflows.

Aritra Sarkar (Delft University of Technology, QuTech), Jul 2021 – Aug 2021 | Remote

QIntern (*Reinforcement Learning Agent for Quantum Foundations*)

- The project involved using a universal AGI RL framework for simple experiments in QIT. The agents interact with the environment to model the quantum observables.
- Received Third Team Award based on best project during the internship.

Artificial Brain, *Quantum Machine Learning Intern* Feb 2021 – May 2021 | Remote

- Developing Quantum Machine Learning models for real-world use cases.
- Exploring the various applications of Quantum Machine Learning in fields like drug discovery, solving many body problems, etc.
- Implementing Quantum Algorithms for various applications including classification, reinforcement learning and protein-folding.

SKILLS

Machine Learning: Linear Regression, Logistic Regression, Decision Trees, Random Forests, Ensemble Techniques - Adaboost, XGBoost, etc., Clustering - Kmeans, DBScan, etc.

Deep Learning: Activation Functions, Optimizers, Loss Functions, Artificial Neural Network, Convolution Neural Network, Recurrent Neural Network, Transfer Learning

Natural Language Processing/LLMs: LSTM, Bi-Directional LSTM, Encoder and Decoder, Attention Models, BERT, Transformers, GPT-3.5-turbo, GPT-4, Gemini, Claude

Languages and Frameworks: Python, SQL, PyTorch, Langchain

Tools and Technologies: Linux, Git, SQL Server, PostgreSQL, MongoDB Atlas, Streamlit, Hugging Face Spaces, NumPy and other Machine Learning Libraries, OpenText InfoArchive

Cloud Platforms: Azure Cloud Services, Google Cloud Platform

CERTIFICATES & BADGES

Microsoft Certified: Azure Fundamentals
Issued by Microsoft

ML Practitioner Certificate
Issued by Dataiku

Hands On Essentials - Data Warehouse
Issued by Snowflake

MongoDB SI Associate Certificate
Issued by MongoDB

Dataiku Core Designer Certificate
Issued by Dataiku

Dataiku Advanced Designer Certificate
Issued by Dataiku

Generative AI for Developers
Issued by Google Cloud Platform

EDUCATION

University of Kalyani, *M.Sc. in Data Science* 2019 – 2021 | Kalyani
- Elective - Nanoscience and Nanotechnology

Raja Peary Mohan College (University of Calcutta), *B.Sc. in Mathematics* 2015 – 2019 | Uttarpara
- Pass Subjects - Physics, Computer Science

CERTIFICATE COURSES

Qiskit Global Summer School, *IBM* Jul 2021 | Online
- Focus on Quantum Machine Learning

Introduction to Quantum Computing, *IBM and The Coding School* Oct 2020 – May 2021 | Online
- Focus on basics of Quantum Computing and Quantum Algorithms

PERSONAL PROJECTS

Quantum Pattern Matching, *Paper Implementation* Apr 2021
- Successfully implemented the paper "Quantum Pattern Matching" by P. Mateus and Y. Omar. Paper Link - https://arxiv.org/abs/quant-ph/0508237

AI Learns to play Flappy Bird, *Self-Learning Project* Nov 2020
- Developed the Flappy Bird game on Python using PyGame that learns to play itself implementing Genetic Algorithm with NEAT-Python