ARPAN GHOSH

AI ENGINEER

https://github.com/arpx17
linkedin.com/in/arpx22
iamarpanghosh@gmail.com
+91-9883368452

Objective:

AI Engineer with a strong background in building AI-driven systems and agents for financial investment decision-making, including developing AI agents to generate precise financial investment memos and leveraging advanced Multimodal Models to handle tables, images, and text. Experienced in implementing and optimizing Retrieval-Augmented Generation (RAG) systems with techniques like Cross-Encoders and Multi-Vector Re-rankers. Adept at designing data pipelines, automating document workflows, and integrating cutting-edge technologies like GPT-4, OpenSearch, and LangChain. Passionate about leveraging Explainable AI and Generative AI to create scalable, innovative solutions that enhance data-driven decision-making and streamline complex workflows.

Skills:

- **Language and Services:** Python, SQL, Snowflake, Azure, AWS.
- **Concepts and Theories:** Generative AI (Gen AI), Prompt Engineering, RAG, RAG Re-rankers, Large Language Models (LLM), Multimodal Models, Multimodal Embedding Models, Fine Tuning (PEFT, LORA, QLORA), Transformers, GAN, Diffusion Models, ASR Models, TTS Models, Vector Databases, Graph Databases, Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), Data Analytics, Physics, Mathematics, Statistics.
- **Tools and Modules:**
  - LangChain, LangGraph, LangServe, LangSmith, Llama-Index, Chroma, FAISS, Pinecone, OpenSearch, Azure AI Search, Azure Document Intelligence, Azure ML Studio, AWS Sagemaker, JupyterLab, OpenAI, Llama, Mistral, GPT , Claude , Gemini , Vertex , Bedrock , Whisper, ElevenLabs , AWS Polly, Azure Speech Studio, OpenCLIP, Colbert, Cohere, Neo4J (Knowledge Graph), Chroma, FAISS, Pinecone, OpenSearch.
  - PyTorch, TensorFlow, SciKit-Learn, Pandas, Streamlit, FastAPI Matplotlib, Seaborn, Plotly, Huggingface, Lime.
  - Azure VM, Azure Blob, AWS EC2, S3, Docker, Git, Linux/Unix, WSL, VS code, Vim

Education:

- M.Sc. in Data Science — 2020 - 2022
  Maulana Abul Kalam Azad University of Technology (WBUT) — CGPA: 9.48
- B.Sc. Physics (H) — 2016 - 2020
  Surendranath College, University of Calcutta — PCT: 66 %

Professional Experience:

GrowthSphereAI **Sept 2024 – Present**

Role: AI Engineer

Currently working at a product-based company building an AI-powered SaaS platform aimed at transforming investment decision-making through Generative AI and Explainable AI. The platform leverages advanced LLMs and agent-based architectures to automate time-consuming analysis, reduce cognitive and confirmation biases, surface implicit assumptions, and identify investment patterns for continuous strategic improvement.

➤ Custom Template Builder Agent:

- Designed and implemented a LangGraph-based AI agent system to automate custom memo generation from multiple sample PDFs.
- Built a pipeline that classifies and clusters similar memos, minimizing template redundancy by aggregating stylistic patterns.
- Automatically generates structured JSON templates that reflect the user's memo style, significantly reducing manual formatting effort.
- The system seamlessly integrates with user-facing applications, enabling consistent memo generation with minimal human intervention.

➤ Image Search & Agentic Infographic Extraction Module:

Led development of an **end-to-end image and table extraction system** tailored for investment documents and reports.

- Used Microsoft Table Transformer models for extracting structured tables and applied OpenCLIP embeddings to index both image and text content in OpenSearch.
- Built an AI agent to classify image types (infographics, charts, tables) and dynamically route them to appropriate extraction logic.
- Integrated reflex-based validation agents to ensure quality and accuracy of the extracted content.
- Developed a multimodal retrieval system where, during search, images are interpreted by an agent that generates structured outputs based on infographic types, which are then passed to LLMs for final response synthesis.
- Enabled unified search across text and image content, improving document retrieval relevance and user experience in a financial decision-making context.

➤ Enhancing RAG with Re-Rankers:

Improved retrieval accuracy by integrating re-rankers on top of the traditional RAG approach. Implemented models such as Cross-Encoders, Multi-Vector Re-Rankers, and T5-based re-rankers to enhance search relevance and response quality.

➤ Contributions to the Auto-QA Module:

Developed a module that converts user queries into OpenSearch DSL queries using an LLM. The LLM identifies relevant filters from the user's query and generates corresponding DSL filter queries. The DSL query is then used to extract relevant chunks, enabling the LLM to generate accurate responses.

➤ Developed a pipeline to scrape data from company websites that investors are interested in, indexing the data into OpenSearch.

➤ Fine-tuned prompts to improve the generation and quality of memos and reports.

➤ Automated document text extraction workflows, indexing data into Vector Indices, and integrating external data sources.

- Collaborated on writing and maintaining test cases for robust validation of AI services.
- Utilized tools such as Docker, LangChain, LangGraph, LangServe, and Python to deliver scalable, high performance AI solutions.

Tata Consultancy Services **Aug 2022 – Sept 2024**

Role: AI/ML Engineer at BFSI A&I Lab

Project Works:

- **Knowledge Graph based RAG for Call Transcript data:**

Utilized automatic speech recognition services from Azure, Deepgram, AWS, and OpenAI Whisper to process call recordings between customers and service executives. Generated and evaluated transcripts for each model, assessing their performance. Created a Neo4j-based knowledge graph to extract and organize key information from transcripts, including customer names, account numbers, investment details, queries, and resolutions. Integrated RAG and knowledge graph outputs into a chatbot, enabling hybrid search capabilities. Developed a system that performed Cypher queries on the knowledge graph and vector-based searches on transcripts, merging results to generate responses using GPT-4. This approach streamlined data retrieval and improved the efficiency and accuracy of customer service interactions, leading to enhanced problem resolution and customer satisfaction.

- **Self-Supervised Recommender for BFSI Clients:**

This is a unique and first in the industry solution developed by TCS AI lab for the banking and insurance clients. The goal was to develop an end to end Self-Supervised Agent which will analyse loan and mortgage application data of banks and generate recommendation for the bank(client) how they can change their loan policy or introduce new loan formats by considering some parameters so that they can get more business from the existing rejected loans or mortgage applications. This will remove the requirement of analysing the data manually using BI tools. This solution was implemented into snowflake platform and sold this solution to more than 5 bank clients of TCS as well as snowflake.

- **RAG Agent using Azure AI Search:**

Developed a RAG agent using Azure AI Search for 3 different clients from BFSI and ESG sector. The goal was to extract information from 1000+ pdf documents which includes images, handwritten texts, some minor infographics and tables. Integrated Azure Document Intelligence, Computer Vision with the AI Search service and used GPT-4 model to get query responses. Used Streamlit to create a basic interactive UI.

Interests and Activities:

- Completed 3rd year at Prayag Sangeet Samiti in Vocal
- Completed 3rd year in painting from Bangiya Sangeet Parishad
- Regularly Practice and play Badminton in an Academy