AI-Powered Domain Expert Chatbot

Intern Name: Kaushal Sahu

Internship Duration: 2/7/2025 – 5/5/2025

Organization: NullClass

1. Introduction

This report outlines the internship project involving the development of an AI-powered chatbot trained on the arXiv dataset. The chatbot is designed to act as a domain expert, capable of answering complex queries and give follow-up answers in the field of computer science(A chosen subset of the dataset) using advanced NLP techniques.

2. Background

With the rapid expansion of scientific literature, accessing and understanding research has become challenging. The goal was to create a dynamic chatbot that extracts, summarizes, and explains research papers using open-source models like LLaMA and technologies such as FAISS, Streamlit, and sentence transformers.

3. Learning Objectives

- Understand how to process and structure large research datasets (arXiv).
- Learn to integrate open-source LLMs like llama for intelligent responses.
- Build a Streamlit-based UI for real-time interactions.
- Expand the chatbot's knowledge dynamically.

4. Activities and Tasks

- Collected and Preprocessed the arXiv dataset (JSON format).
- Extracted abstracts and metadata for vectorization.
- Generated embeddings using SentenceTransformer.
- Created and queried a FAISS index for document retrieval.
- Used a local LLaMA model to generate answers using relevant paper excerpts.
- Developed a clean UI with Streamlit for interaction like Chat GPT.
- Added support for future expansion and updates.

5. Skills and Competencies

- Python (file handling, NLP, APIs)
- Natural Language Processing (SentenceTransformer, summarization)
- Large Language Models (LLaMA, inference logic)
- Vector Search (FAISS)
- UI/UX using Streamlit
- Data preprocessing and JSON parsing
- Problem-solving and debugging

6. Feedback and Evidence

The chatbot was successfully tested using complex queries from actual arXiv papers and follow-up questions including which paper the model have used to give the answers the queries asked. It handled the data very well and answers very smartly.

A sample video of conversation with this chatbot have been shared with the file. File named - “Sample_modelTest.mp4”.

7. Challenges and Solutions

1. ***Understanding arXiv’s JSON Structure of all the research paper. It was tricky to extract the information from the JSON file***

SOL- Wrote a parses to extract titles, abstracts. It become easier once the file structure was decoded and the code extracted the titles and the abstracts of the papers in an different file. As a subset

2. ***Managing large vector data***

SOL- It was hard to process the whole 3.5lakh paper and store it. I tried using online RAM and GPU but the time taken to embed was so much. Finally I decided to embed the file in batches(for faster and secure approach and also I will have a track), and trained the model with 8000 papers as an base chatbot approach.

3. ***LLaMA local setup confusion***

SOL- The model used in pervious task was only useful for that specific Task. To make the llama train on the be smart enough to give a follow-up answer I have to use the LLaMA3 model with the custom data.

4. ***Making streamlit UI and applying follow-up approach***

SOL- To make the chatbot show the follow-up queries with the history written above(a UI like ChatGPT). I had to take a major help from ChatGPT(as it was not taught to me in the training session). But after some bug fixing and coordinating with GPT. UI and the structure was finally working and made possible.

8. Outcomes and Impact

- Built a fully functional, topic-specific research assistant.
- Learned full-stack AI project workflow from dataset to deployment.
- Learned the use of different AI APIs and there coding differences.
- Demonstrated real-world application of AI for knowledge automation.

9. Conclusion

This internship project has been a transformative learning experience. Working with the arXiv dataset allowed me to engage with real-world, high-quality scientific content, which helped me sharpen my data wrangling and NLP understanding skills. It taught me how to structure AI pipelines, and gave me hands-on experience in solving complex problems through creative code. The chatbot can be scaled further and applied across multiple research domains.