**Bx-BOT**

**Abstract**

In this project, we develop an intelligent chatbot system by leveraging data scraped from a website, structured and stored in JSON format. The core objective is to enhance the chatbot's response accuracy and contextual relevance by integrating this data into a vector database. By employing a Retrieval-Augmented Generation (RAG) approach, we utilize the LLaMA 2 language model to retrieve the most relevant information from the vector database, which is then used to generate accurate and context-aware responses. The integration of a vector database with the LLaMA 2 model allows the chatbot to effectively manage and query vast amounts of unstructured data, resulting in a more powerful and efficient system. The outcome of this project is a robust chatbot that can interact with users more naturally, providing precise answers by combining the strengths of retrieval-based and generation-based methods.

**Steps To Be Followed And Working of Process**

**Data Extraction**: Using web scraping tools like **BeautifulSoup** to extract the required information from the **BIMDeX** website.

**Data Cleaning**: Remove any unnecessary or noisy data (e.g., HTML tags, JavaScript snippets) from the scraped content.

**Text Normalization**: Convert text to a consistent format, such as lowercase, and remove punctuation, special characters, or stop words

**Text Normalization**: Convert text to a consistent format, such as lowercase, and remove punctuation, special characters, or stop words

**Text Embedding**: Convert the text data into vector representations using pre-trained models like BERT, **Sentence Transformers**, or custom embeddings

**Storing Vectors**: Store the generated vectors in a vector database (e.g., Pinecone, FAISS, Milvus) for efficient retrieval. Here we are using **MILVUS** database for open and free source and capable for managing high dimension vectors

**Model Setup**: Setting up the **LLaMA2 (Meta)** model in your environment. This involves configuring it to interact with the vector database.

**Query Processing**: When a user inputs a query, convert it into a vector using the same embedding model.

**Vector Retrieval**: Retrieving the most relevant vectors from the database based on similarity to the query vector.

**Contextual Response Generation:** Feed the retrieved vectors into the **LLaMA2**  model to generate a response that is both accurate ,relevant

**WorkFlow Visualisation**

