**Report on Retrieval Augmented Generative Chatbot for Urdu Stories**

**Introduction:** This report details the implementation of a Retrieval Augmented Generative Chatbot for Urdu Stories, showcasing a sophisticated Conversational AI system. The system incorporates various natural language processing (NLP) techniques, including text preprocessing, embeddings, translation, and interaction with a pre-trained language model.

**1. Data Preprocessing:**

1.1 **Urdu Stories Upload and Cleaning:**

* The initial step involves uploading a dataset of Urdu stories and cleaning the data. The stories are read from a CSV file, split into sentences, and cleaned by removing special characters, stopwords, and other unnecessary words.

1.2 **Embeddings Creation:**

* Sentence embeddings are generated using the SentenceTransformer library. The 'all-MiniLM-L6-v2' pre-trained model is employed to create numerical representations of the cleaned Urdu sentences.

**2. Vector Database with Faiss:**

2.1 **Faiss Indexing:**

* The Faiss library is utilized to create a vector database for efficient similarity search and retrieval. The Inverted File with Flat (IVF) method is employed, and the trained index is saved to a file for future use.

**3. Audio Transcription:**

3.1 **Automatic Speech Recognition (ASR):**

* Audio files are transcribed using the Whisper ASR model, specifically trained for Urdu. The Hugging Face Transformers library is utilized for ASR, simplifying the process of converting spoken language into text.

**4. User Text Embedding:**

4.1 **Text Embedding with SentenceTransformer:**

* User input in Urdu is preprocessed by removing stopwords and irrelevant words. The SentenceTransformer model is then used to create a numerical embedding for the user's input, facilitating various NLP tasks.

**5. Finding Relevant Sentences:**

5.1 **Cosine Similarity Search:**

* Relevant sentences from the dataset are identified based on cosine similarity with the user's input embedding. The Faiss index is queried to find the top k most similar sentences, enhancing the system's ability to provide contextually relevant information.

**6. Translation with MarianMT:**

6.1 **Urdu to English Translation:**

* The Hugging Face Transformers library is employed to translate both the user's input and the selected relevant sentences from Urdu to English. The MarianMT model ('Helsinki-NLP/opus-mt-urdu-en') is used for this task, broadening the system's language capabilities.

6.2 **English to Urdu Translation:**

* The translated English sentences are then converted back to Urdu using the reverse translation process. The MarianMT model ('Helsinki-NLP/opus-mt-en-urdu') is utilized for this purpose, ensuring a seamless transition between languages.

**7. Conversational Interaction with Replicate:**

7.1 **Answer Generation with Llama2:**

* The Replicate platform is leveraged for conversational interaction. The Llama2 model is used to generate answers based on the provided context and user questions, enhancing the system's ability to engage in dynamic conversations.

**Conclusion:** The implemented Retrieval Augmented Generative Chatbot for Urdu Stories showcases the integration of various NLP techniques, providing a sophisticated system capable of understanding and responding to user input in Urdu. The system facilitates cross-language interactions and delivers relevant information based on contextual understanding.