**Knowledge-Retrieval Chatbot (RAG-based Q&A System)**

**Overview:**

The main goals of this project were to integrate optimized LLMs, deploy the program using Streamlit, and develop a chatbot through Retrieval-Augmented Generation (RAG). The goal was to create an AI-powered assistant that could extract pertinent data from a dataset and then produce insightful answers. Stable Diffusion was also investigated as a component of this multi-modal AI model for improving user experience.

**Specifics of Implementation:**

1. Pipeline for Retrieval-Augmented Generation (RAG)

* For effective document retrieval, the chatbot was built to use FAISS  to extract pertinent information from a dataset.
* A already trained LLM (such as Flan-T5) was adjusted to produce answers according to the data that was recovered.
* By using retrieval strategies to improve response relevance, the system guaranteed factual correctness.

2. Streamlit Chatbot Development

* An interacting web-based chatbot interface was developed using Streamlit.
* The chatbot was made to take user inquiries, find relevant information, and produce responses that resembled those of a human.
* Before the application was ready for cloud deployment, it underwent local testing.

3. Using Cloud Services and ngrok for deployment

* Initially, ngrok was used to deploy the chatbot locally for public access.
* Fixing up the authtoken correctly fixed the ngrok authentication problems.
* In order to facilitate remote access and scalability, cloud deployment options like Hugging Face Spaces & Heroku were investigated.

**Difficulties Met:**

* Problems with Ngrok Authentication: You need to create an account that has been verified and set up the authtoken correctly.
* Errors in Port Configuration: When Streamlit was launched with ngrok, the first errors were caused by improper port configuration.
* LLM Fine-Tuning Complexity: It took several iterations to modify the model in order to increase response accuracy.
* Deployment Troubleshooting: Troubleshooting compatibility concerns was necessary to ensure smooth execution on cloud platforms.

Related Screenshots:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a chatbot

AI-generated content may be incorrect.

A screenshot of a chatbot

AI-generated content may be incorrect.

A screenshot of a chatbot

AI-generated content may be incorrect.