**Group 14**

Gagandeep Singh - 100897670

Ashwin Philip - 100906431

Preeja Vimala Antony -100899911

**AI Product Design Report for Language Model-Powered Q&A Bot**

**Introduction**

In this AI product design report, we present the development of a Language Model (LLM)-powered Question and Answer (Q&A) Bot. The aim of this project is to leverage the capabilities of advanced language models to create an interactive and user-friendly AI application that can provide answers to user queries based on a given input text document. The Q&A Bot utilizes the Gradio.io library for the front-end interface and integrates with Hugging Face's transformers for the language model implementation.

**Problem Identification**

The primary problem that our AI product addresses is the need for efficient and accurate information retrieval from large text documents. In many real-world scenarios, users often struggle to extract relevant information from extensive textual sources, such as research papers, articles, or technical documents. Our AI-powered Q&A Bot aims to streamline this process by allowing users to input a text document and interactively ask questions to retrieve specific information from within the document.

**Language Model Selection**

For this project, we have selected the "google/flan-t5-large" model, which is a powerful text-to-text generation model provided by Hugging Face. This model demonstrates robust performance in generating coherent responses to user queries and can handle a wide range of textual input formats. The choice of this model aligns well with our goal of developing an effective Q&A Bot.

**AI Product Design**

Our AI product consists of two main components: the backend, which includes the language model and the index, and the front-end application, which enables user interaction with the Q&A Bot.

**Backend:**

* Custom LLM Class (customLLM): We engineer a specialized LLM class named customLLM, derived from the "google/flan-t5-large" model. This class is dedicated to orchestrating text generation tasks and is the engine behind our Q&A Bot's response generation.
* The GPTVectorStoreIndex is used to query for a response based on the user input.

**Front-end Application:**

* Gradio.io Integration: We leverage Gradio.io to craft an interactive and intuitive front-end interface, optimizing user engagement with the Q&A Bot.
* Input Text Document Tab:

Seamless interaction commences in this tab, where users seamlessly input their desired text document. The "Build the Bot!!!" button initiates the index-building process, a pivotal step ensuring the Q&A Bot's comprehensive grasp of the document's content.

* Knowledge Bot Tab:

Users delve into the heart of AI-powered interaction through the Knowledge Bot Tab. Through an engaging chat interface, users engage in dynamic conversations with the Q&A Bot. This real-time exchange allows users to pose questions, and in response, the bot draws upon the meticulously crafted index to furnish accurate and insightful answers.

Our front-end application acts as the bridge between users and technology, enhancing accessibility and ensuring a user-friendly experience. By blending cutting-edge innovation with a sleek design, the application effectively facilitates seamless interaction, making information retrieval an intuitive and rewarding endeavor.

**Technical Implementation**

Our AI product's technical implementation involves a series of steps, each contributing to the seamless integration of a Language Model (LLM) into an interactive and user-friendly Question and Answer (Q&A) Bot.

1. Custom LLM Class (customLLM)

We start by creating a custom LLM class, customLLM, which serves as the backbone of our Q&A Bot's language processing capabilities. This class is built upon the foundation of the "google/flan-t5-large" model, a powerful and versatile language model. The customLLM class encapsulates the following key methods:

* \_call: This method leverages the model pipeline to generate textual responses based on user prompts. It accepts user input and utilizes the model's language generation capabilities to craft meaningful and relevant answers.
* \_identifying\_params: This method returns essential information about the model, such as its name and characteristics, providing insights into the underlying language processing capabilities.
* \_llm\_type: This method defines the type of Language Model being used, in this case, categorizing it as a "custom" LLM.

2. Building the Index

To enable efficient querying and retrieval of information, The GPTVectorStoreIndex is used to create an index that maps user inputs to chatbot-generated responses.​ It's then used to query this index and retrieve relevant responses based on the user's inputs.​The index provides a mechanism for efficient retrieval of responses associated with user inputs, making the interaction with the chatbot more responsive and contextually relevant.​

3. Front-end Interaction using Gradio.io

We create an interactive and intuitive front-end interface for users to engage with the Q&A Bot. Gradio.io, a user-friendly library, empowers us to design an accessible interface where users can seamlessly interact with the Q&A Bot.

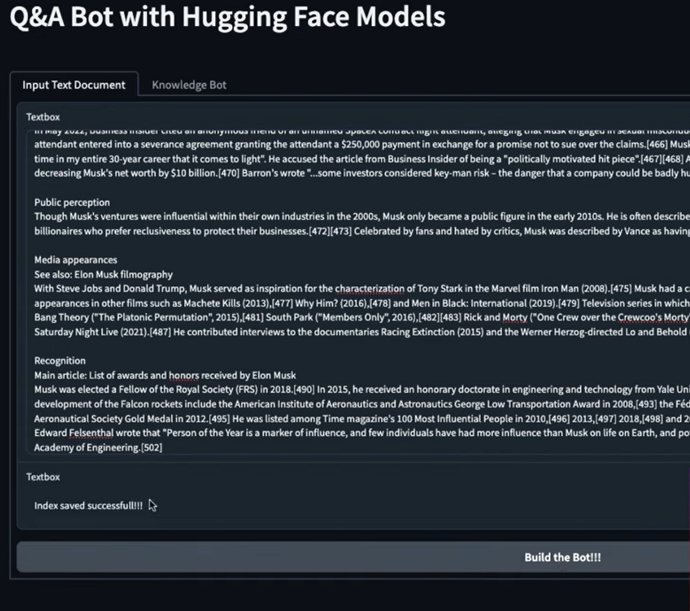
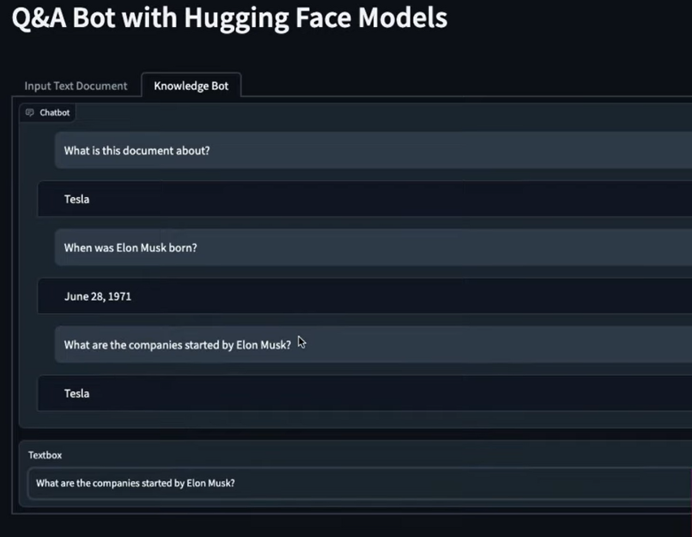
* Input Text Document: Users can input a text document of their choice. Upon submitting the document, they can initiate the process of building the Q&A Bot's index by clicking the "Build the Bot!!!" button.
* Chat Interface: The Q&A Bot's chat interface enables users to pose questions and receive responses based on the generated index. Users input their queries through the chat interface, and the Q&A Bot leverages the index to generate relevant and accurate answers.

4. Interaction and Response Generation in Chat

The Q&A Bot processes user requests using the index in the chat interface. The Bot recognises key areas of the document and generates responses using the custom LLM (customLLM). Users can see the responses in real time, which improves the interactivity and reactivity of the Q&A experience.

In summary, our Language Model-powered Q&A Bot's technical implementation entails the building of a custom LLM class, the creation of an index for quick data retrieval, the usage of Gradio.io for user interaction, and the integration of chat-based interaction with the Q&A Bot. This seamless integration of components guarantees that consumers can effectively interact with the Q&A Bot, obtain accurate responses, and have a better interaction experience.

**Screenshots:**



**Limitations and Ethical Considerations**

**Limitations**

* Limited Context

The Q&A Bot, while adept at generating responses based on the input text document, operates within the confines of the provided context. Consequently, the accuracy and relevance of its answers are closely tied to the information available within that context. Queries that deviate significantly from the context may result in less accurate or even irrelevant responses. It's important for users to be mindful of the context when interacting with the Q&A Bot to ensure optimal results.

* Complex and Ambiguous Queries

The Q&A Bot may encounter challenges when faced with complex or ambiguous queries that require nuanced understanding or contextual interpretation. In such cases, the generated responses might not fully capture the intended meaning of the query, potentially leading to inaccuracies or incomplete answers.

**Ethical Considerations**

* Bias and Fairness

The responses generated by the Q&A Bot are influenced by the language model's training data, which may contain inherent biases present in societal, cultural, and historical contexts. As a result, the Q&A Bot's answers could inadvertently reflect or amplify biases present in the data. We are committed to ongoing efforts to mitigate and reduce biases, but it's essential to be aware of this potential limitation.

* User Responsibility

Users of the Q&A Bot bear a certain level of responsibility in interpreting and utilizing the information provided. While the Q&A Bot aims to assist and provide accurate answers, users should exercise critical thinking and verify information, especially for important decisions or sensitive topics.

* Privacy and Data Usage

Interactions with the Q&A Bot involve sharing input text documents, which may contain sensitive or proprietary information. Ensuring the privacy and secure handling of user data is a paramount concern. We take measures to protect user data and adhere to data protection regulations, but users should exercise caution and avoid sharing confidential or sensitive information.

**Conclusion**

While our AI product offers valuable benefits in facilitating information retrieval and enhancing user interactions with textual data, it's important to acknowledge and address these limitations and ethical considerations. By being transparent about potential challenges and risks, we aim to foster responsible and informed usage of our Q&A Bot and continuously work towards improving its performance, fairness, and overall user experience.