Brance AI Applied Researcher Role Task

Name: Manish Agarwal

LinkedIn Profile: https://www.linkedin.com/in/manish-coder/

Date Challenge Received: 06 July 2023

Date Solution Delivered: 08 July 2023

1. Problem Statement

The problem requires building a chatbot using Retrieval Augmented Generation (RAG) that can answer user questions based on a provided knowledge document. The RAG module should consist of a retrieval phase to find the relevant context from the knowledge document and a generation phase powered by a Language Model (LLM) to personalize the answer using the retrieved knowledge. The main features required are answering user questions from the document and preventing hallucination. There is also an optional bonus feature of evaluating the relevance of the answers using a metric.

2. Approach

To solve the problem, I would follow these steps:

* Understand the structure and format of the knowledge document.
* Preprocess the document to prepare it for retrieval and generation.
* Create a dense vector space of the preprocess document using the FAISS Library
* Using Lang chain for the chatbot using the retrieval and OpenAI library where we Integrate a Language Model (e.g., GPT-3) to generate personalized answers using the retrieved context.
* Implement measures to prevent hallucination in the generated answers. (Prompt engineering)
* Optionally, develop a metric to evaluate the relevance of the answers.

Assumptions:

* The knowledge document is in a readable format and can be preprocessed easily.
* Sufficient computational resources are available to train and deploy the models.
* An LLM orchestrator like Lang chain or a similar framework can be utilized for the RAG module.

3. Solution

The solution would involve implementing the RAG module using a combination of retrieval and generation techniques. The retrieval phase would use the knowledge document to find relevant context based on user questions. The generation phase, powered by an LLM, would take the retrieved context, and generate personalized answers to the user questions.

A diagram of a chatbot

Description automatically generated

The retrieval module retrieves relevant context from the knowledge document based on the user's question. The generation module takes the retrieved context as input and generates personalized answers using an LLM.

4. Future Scope

To improve the solution, some possible areas to consider are:

* Fine-tuning the retrieval model: The retrieval phase can be further optimized by fine-tuning the model using techniques like BERT or dense retrieval approaches.
* Training a more advanced generation model: The generation phase can benefit from using more advanced language models like GPT-4 or GPT-Neo, which may provide better-quality responses.
* Enhanced evaluation metric: Develop a robust evaluation metric to measure the relevance and quality of the generated answers, considering both the user's question and the retrieved context.
* User interface improvement: Create a user-friendly interface to interact with the chatbot, allowing users to easily input questions and receive answers in a conversational manner.
* Deployment and scalability: Optimize the solution for deployment in a production environment, considering factors like performance, scalability, and resource utilization.

These improvements can enhance the chatbot's accuracy, usability, and overall performance.