Task Report

Fine-Tuning with Retrieval-Augmented Generation (RAG) Chatbot

0. Task Location:

1. **Github:** https://github.com/I-Quit-Bro/flan-t5-large-finetuned
2. **Google Drive:** https://drive.google.com/drive/folders/1tgVdMI-uIUQLW2xcxx9y019ww6-Ao_Tp?usp=sharing
3. Google Colab: https://colab.research.google.com/drive/1YLi0gPU0l85qkfe0i1IiKAEVGYR81hVP?usp=sharing

1. Project Overview

This project is about creating and improving a medical chatbot that uses Retrieval-Augmented Generation (RAG). RAG helps the chatbot give better answers by searching the knowledge base for the most relevant information, ranking the results based on their importance, and then combining that information with its language skills to create a clear and accurate response. The chatbot is fine-tuned using a large language model called FLAN-T5, paired with a model for finding information efficiently, called `all-MiniLM-L6-v2`.

2. How does the model work?

The chatbot's system includes:

- **Main Language Model:** `google/flan-t5-large`. This pre-trained model helps the chatbot understand and respond to questions. It’s especially good for tasks like answering questions and summarizing information.
- **Embedding Model:** `all-MiniLM-L6-v2`. This model finds related information in the knowledge base by creating compact, meaningful text representations.
- **RAG Framework:** The chatbot uses RAG to first find relevant information in the knowledge base, then combine that information with its language skills to generate complete and accurate answers.

3. Knowledge Base

The chatbot uses a knowledge base with information on different medical topics, such as:

- **Heart Conditions:** High blood pressure and coronary artery disease.
- **Diabetes and Metabolism Issues:** Type 2 diabetes and insulin resistance.
- **Lung Diseases:** Asthma and chronic obstructive pulmonary disease (COPD).
- **Brain Disorders:** Alzheimer’s disease and Parkinson’s disease.
- **Cancer:** Breast and lung cancer, along with their treatments.
- **Mental Health:** Depression and anxiety.
- **Infectious Diseases:** Tuberculosis and COVID-19.
- **Bone and Joint Problems:** Osteoarthritis and fractures.
- **Children’s Health:** Pediatric asthma and common viral infections.
- **Digestive Problems:** IBS and GERD.

This knowledge base ensures the chatbot can provide reliable answers based on the topics above.

4. Key Design Decisions

Several important factors were considered to make the chatbot work well:

1. **Choosing the Right Models:**
    - The FLAN-T5 model was chosen for its advanced understanding of language and adaptability to medical questions.
    - The embedding model, `all-MiniLM-L6-v2`, was selected for its ability to quickly and accurately find related information.
2. **Organizing the Knowledge Base:**
    - The chatbot's knowledge base contains curated information on many medical conditions and treatments.
    - It is updated regularly to include the latest medical guidelines and research.
3. **Using RAG Effectively:**
    - RAG helps the chatbot combine information retrieval and language generation, ensuring responses are accurate and detailed. For example, if a user asks about treatments for asthma, RAG will search the knowledge base for relevant documents about asthma treatments, rank them based on importance, and then use this information to generate a clear and concise answer.
4. **Scaling the System:**
    - The system can grow by adding more medical documents, ensuring it stays useful as more topics are covered.
5. **Ensuring Accuracy:**
    - By fine-tuning the chatbot and using RAG, responses are not only detailed but also factually correct.

5. Results

- The chatbot consistently generated clear, accurate responses to various medical questions.
- It handled queries quickly, with minimal delay.
- New documents could be added to its knowledge base without affecting performance.
- Its ability to find the right information made it highly reliable for users.

6. Practical Results (responses:

Prompt: A 50-year-old male with obesity presents with chronic knee pain, which is worse with activity and relieved by rest. Physical examination reveals knee joint tenderness and crepitus with movement. What is the likely diagnosis and the next steps?

Chatbot: Osteoarthritis is a degenerative joint disease that causes pain, stiffness, and swelling, particularly in weight-bearing joints such as the knees and hips.

Prompt: A 26-year-old man has been experiencing persistent restlessness, impulsivity, and difficulty concentrating both in his professional and personal life. He also mentions a history of similar problems during childhood. What could be the potential causes?

Chatbot: ADHD is characterized by symptoms like restlessness, impulsivity, and difficulty concentrating, often beginning in childhood. It is caused by the loss of dopamine-producing neurons.

Prompt: A 45-year-old woman fell onto her outstretched hand while jogging. She now has pain and swelling in her wrist and difficulty moving it. What is the likely diagnosis and what should be the next step?

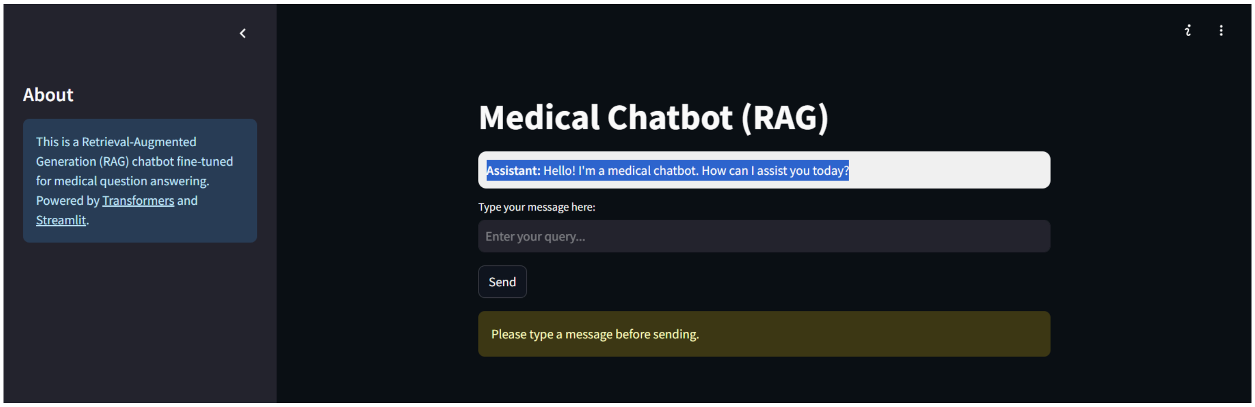
Chatbot: A wrist sprain or fracture is a common injury that can result in pain, swelling, and difficulty moving the wrist. Common injuries include scaphoid fractures, distal radius fractures, or wrist sprains. These injuries can be diagnosed with X-rays or other imaging techniques. Osteoarthritis is a degenerative joint disease that causes pain, stiffness, and swelling, particularly in weight-bearing joints such as the knees and hips.

7. Future Improvements

To improve the chatbot, we plan to:

1. **Create an Functioning User Interface (UI):**
    - Develop an easy-to-use interface for better user interaction.
    - Add features like conversation history, bookmarks, and suggestions for follow-up questions.

Example UI (under process):



About

This is a Retrieval-Augmented Generation (RAG) chatbot fine-tuned for medical question answering. Powered by Transformers and Streamlit.

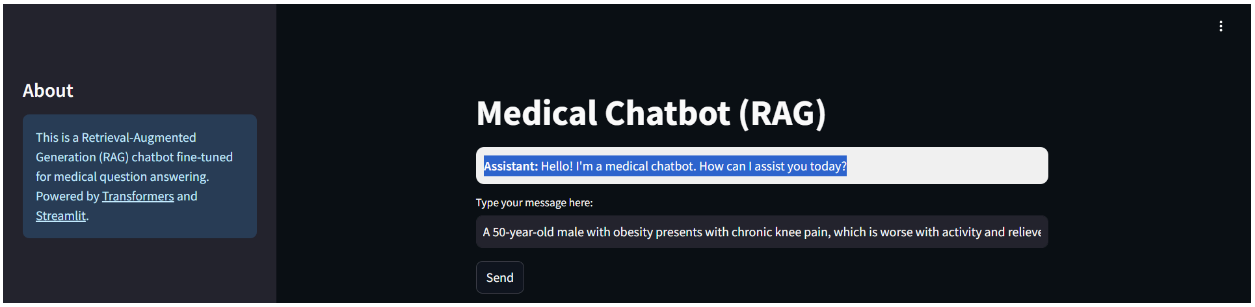
Medical Chatbot (RAG)

Assistant: Hello! I'm a medical chatbot. How can I assist you today?

Type your message here:

Enter your query...

Please type a message before sending.



About

This is a Retrieval-Augmented Generation (RAG) chatbot fine-tuned for medical question answering. Powered by Transformers and Streamlit.

Medical Chatbot (RAG)

Assistant: Hello! I'm a medical chatbot. How can I assist you today?

Type your message here:

A 50-year-old male with obesity presents with chronic knee pain, which is worse with activity and relieve

2. **Allow Longer Conversations:**
    - Enhance the chatbot's ability to handle longer and more detailed discussions without losing track of the context.
3. **Update the Knowledge Base Automatically:**
    - Use automated tools like web crawlers, natural language processing pipelines, and API integrations to keep the knowledge base up to date

with the latest medical research. These tools can automatically extract, validate, and add new information from trusted medical sources.

8. References

1. "FLAN: Few-Shot Learning with Natural Language Processing," Google Research.
2. "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," Microsoft Research.
3. RAG Implementation Documentation, Hugging Face.
4. Medical references collected from peer-reviewed journals and trusted online resources.