FinalProject-Group3

Project Proposal for DoctorGPT: An AI-Based Medical Q&A Chatbot

1. Problem Selection

The goal of this project is to build a DoctorGPT chatbot, a specialized Question-and-Answer (Q&A) system focusing on medical inquiries. There is a high demand for AI-powered solutions that can assist patients with common medical questions, especially in areas with limited healthcare access or long wait times for consultations. By creating a specialized chatbot, we aim to improve preliminary healthcare guidance and empower users with reliable information before consulting with a medical professional.

2. Dataset

We will use the "AI Medical Chatbot" dataset from Hugging Face, which includes three columns: Description, Patient Question, and Doctor Answer. This dataset provides a solid foundation with realistic doctor-patient dialogues, offering ample training data for our Q&A model.

3. NLP Methods

For this project, we will employ a combination of classical and customized NLP methods:

1. **Topic Modeling**: We will start by conducting topic modelling on the Description column to identify relevant medical topics. Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) will be considered to group similar topics.
2. **Model Training**: We will fine tune model(LLM) using the Q&A pairs in the Hugging Face dataset to understand and generate responses that align with the format and accuracy expected in medical Q&A contexts.
3. **RAG**: We will create a RAG on the basis of topic modelling leveraging Wikipedia and medical texts.

4. Packages

The primary packages will include:

- **Hugging Face Transformers**: For model deployment, specifically fine-tuning transformer models for medical Q&A.
- **Scikit-learn**: For model evaluation metrics and additional preprocessing.
- **Wikipedia API**: For scraping supplementary topic data.
- **PyTorch/TensorFlow**: For deep learning model training and adjustments.

These packages offer efficient, robust libraries for NLP tasks and deep learning, ideal for processing large datasets and fine-tuning models for specific medical queries.

5. NLP Tasks

The project will focus on the following NLP tasks:

- **Topic Modeling**: Extract meaningful clusters from medical descriptions to build a knowledge base.

- **Question Answering**: Fine-tune the model to respond accurately to medical questions based on learned patterns.
- **RAG-Component**: To able to refer medical texts and Wikipedia for extra knowledge.

6. Performance Evaluation

We will judge the model's performance using:

- **Human Evaluation**: For assessing the relevance and readability of responses.

7. Project Timeline

- **Week 1**: Data acquisition and exploration, topic modeling on the description column.
- **Week 2**: Wikipedia data scraping and researching on models to Finetune
- **Week 3**: Fine-tuning the model.
- **Week 4**: Building RAG.
- **Week 5**: Performance evaluation, metric analysis, and final tuning.
- **Week 6**: Documentation, project review, and final adjustments for presentation.

This structured approach will allow us to create an effective DoctorGPT that provides accurate and relevant medical answers while ensuring high-quality user interactions.