Optimizing Question-Answering (QA) Performance in Large Language Models(LLMs): A Study in Parameter-Efficient Fine-Tuning (PEFT), Evaluating Retrieval-Augmented Generation (RAG) and Retrieval-Augmented Fine-Tuning (RAFT) Systems for Domain-Specific QAs

Project Overview

This project evaluates Fine-Tuned, Retrieval-Augmented Generation (RAG) and Retrieval-Augmented Fine-Tuning (RAFT) systems for domain-specific QAs.

Prerequisites

- Create a Hugging Face account
- Generate an access token with write permissions

Setup Instructions

1. **Creation of Q&A datasets**
    i. This section includes this notebook:
        i. QADataset.ipynb
    ii. Run all the cells in the “QADataset.ipynb” file in order to create the Q&A datasets, including HR general, and Advanced medical Questions and response. These files are created in two possible formats such as csv and json.

2. **Dataset Creation**
    i. This section includes this notebook:
        i. NLPProject_Dataset_MedicalQA.ipynb
    ii. Store the access token in the write mode in the secret keys in order to access it for pushing the dataset on the Hugging Face repository. We use this smaller dataset to fine tune our selected LLM.
    iii. Run the first cell to install all the required libraries in order to run this notebook.
    iv. Login to your Hugging Face by using the same access token stored in the colab notebook.
    v. Load the chosen dataset. You can change it to a preferred one.
    vi. Choose a LLM, the default model is Mistarl-7b-v0.2. You can change it in the tokenizer section.
    vii. Run all the cells respectively.

viii. You can change the token limits in “valid_indices” in order to reduce the number of instructions and outputs.
ix. In order to choose the top_k rows, it’s possible to change the k parameter to create a smaller or larger dataset.
x. Push the created dataset to your Hugging Face repository by running the last cell.

3. Performing Domain-Specific Fine-Tuning

i. This section includes this notebook:
    i. Mistral_7B_Instruct_v0_2_finetuning_medicaldb.ipynb
ii. It follows the same procedure as before by running all the cells from top to down.
iii. You can change the base_model variable to use your preferred LLM.
iv. It is possible to alter the configuration of training setups to fine tune the model in your preferred setups.
v. Push the fine-tuned model to your Hugging Face by running the last cells.

4. Querying Baseline and fine-tuned models

i. This section includes these notebooks respectively. All these notebooks follow the same procedure. They are implemented by different configurations and system prompts to improve the generated responses.
    i. NLPProject_Mistral_7B_Intrsuct_v02_Prompting_QA(P1).ipynb
    ii. NLPProject_Mistral_7B_Intrsuct_v02_Prompting_QA(P2).ipynb
    iii. NLPProject_Mistral_7B_Intrsuct_v02_Prompting_QA(P3).ipynb
    iv. NLPProject_Mistral_7B_Intrsuct_v02_Prompting_QA(P4).ipynb
    v. NLPProject_Mistral_7B_Intrsuct_v02_Prompting_QA(P5).ipynb
ii. In this section you need to upload the created dataset for HR general and advanced medical Q&A on the colab notebook before running the notebook cells.
iii. After uploading the mentioned datasets, It follows the same procedure as before by running all the cells from top to down.
iv. You can also change the load_in parameter to load the model in 4 bit or 8 bit precision.
v. It is possible to alter pipeline parameters to generate responses using different setups such as temperature and top_p.
vi. Save the generated csv and json files containing questions and generated answers by the model to evaluate them.

5. RAG System Implementation

i. This section includes these notebooks. The first notebook is implemented by using the base model and the second one uses the fine-tuned version. All these notebooks follow the same procedure, these two models implemented by the same setups and system prompts to generate responses.
    i. NLPProject_Mistral2_7B_RAG.ipynb
    ii. NLPProject_Mistral2_7B_RAG(FT).ipynb
ii. Here, you need to upload the papers, from the papers directory in the zip file in this directory on the colab notebook "./content/papers" before running the notebook cells.
iii. Then you need to upload the created dataset for HR general and advanced medical Q&A on the colab notebook as the same as before.
iv. You can also change the load_in parameter to load the model in 4 bit or 8 bit precision.
v. It is possible to alter pipeline parameters to generate responses using different setups such as temperature and top_p.
vi. Save the generated csv and json files containing questions and generated answers by the model to evaluate them.

6. RAFT Integration and Evaluation

i. This section includes this notebook:
    i. Mistral_7B_Instruct_v0_2_finetuning_medicaldb.ipynb
ii. In this section you need to upload the papers, from the papers directory in the zip file in a local directory on the colab notebook named "papers" before running the notebook cells.
iii. Then, you need to upload the created dataset for HR general and advanced medical Q&A on the colab notebook as the same as before.
iv. Save the generated csv files containing generated answers to evaluate them.

Contributing and Contacts

Shakiba Farjood Fashalami
- shakiba.farjoodfashalami@studenti.unipd.it

Nima Daryabar
- nima.daryabar@studenti.unipd.it

Tobia Pavona
- tobia.pavona@studenti.unipd.it

Marcos Tidball
- marcos.tidball@studenti.unipd.it

Giacomo Ferrante
- giacomo.ferrante@studenti.unipd.it