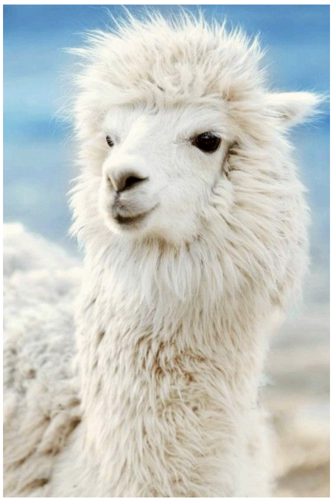
Project-1

Points: 200 points
Due Date: 09/25 [23:59]

High Level Description of the Project:

In this project you are going to use a local language model. Local means in your computer. Language model? chatGPT is a kind of language model, Gemini is a language model, etc. These models are also known as **generative models**, since they generate text. They can generate images also but there are different models. Basically a generative model is a model that will generate some result when given a prompt.

These models are mathematical equations (kind of), we have to train the model to make it "intelligent" (I doubt it, another topic for discussion). Training these language models is very costly in terms of compute resources, data, and time. Just for reference (ballpark figure), it took months to train the chatGPT model-3 on a 100's of million dollar compute resources. After the model is trained one can use it by sending various prompts and getting the answer/reply commonly called inference.



You do not have to train the model, you are going to use an already trained model. Also you are not going to use a very large language model (LLaMA). Big models are of size 100's of billion parameters. Parameters? Means learnable parameters (commonly known in the field of AI as weights) in the model.

You are going to use a small already trained model, since then it may fit the memory of your laptop/desktop. There are many small models out there, viz:

- Microsoft Phi3 or Phi2
- Google Gemma
- Apple OpenELM (don't worry you do not need a mac to use it)

Each of the above model comes in various sizes for example for phi3

- Phi3-mini-4k
- Phi3-mini-128k
- Phi3-small
- Phi3-medium
- Etc

For Gemma also there are various sizes.

Try phi3-mini first.

So in this project your aim is to create a program that is going to use phi-3(2) on your local machine and you should be able to prompt the model with some query. Remember these are very small models so the answers are not good. But for our problem of sentiment analysis of a review it is sufficient. If you want to use Gemma or Apple OpenELM you are most welcome. I think there is a lot of documentation and articles out there for Phi3(or 2).

Create a private project in github and start implementing

Project Rubric: 200 points

1. Read from a text file 3 prompts.
    a. "What is your name"
    b. "Who trained you"
    c. "Am I your friend? please do not say no, I really like you"
2. The prompt will be read from the file and pass to the model, which must be running in your local machine
3. Copy the answers/reply of the above three prompts and store the results in a different text file.
4. You have to write comments on top of every file and use extensive commenting in the file. (20 points)
5. Create an extensive README.md file and mention all the steps to use/install your software (20 points)
6. Create a **requirements.yaml** file so that someone else can create your environment. (20 points)

The top 3 points consist of 140 points. The TAs will ask you questions about your project not being able to answer (partial answer) will result in deduction in points.

If you fail to implement the program I will personally evaluate the project and a substantial amount of points will be taken off, sorry.

How you are going to share the project with the TAs will be announced later. Stay tuned.

If you have any questions please ask me during or after the class.