Report of the Programming Task 1

- The Process- Firstly, I started of working with the data and created all the identity terms into separate lists. This was done so that I could generate all the possibilities for each of the region , religion identity terms.
- Second, then I generated all the possibilities so as to feed them as prompts to the language models.
- I fed them to Google BERT(base-uncased) using Masked Language Model and found out that the model is generally trying to be unbiased by using "they", but if it wants to mention a gender it is biased towards "females".

Report of Programming Task 2

- I started with the same thought of feeding all the possibilities to the LLMs, but I was unaware about the rate limit on the API calls, thus wasting my valuable resources.(used ChatGPT 4o, Google Gemini 1.5 pro).
- Then I manually fed the prompts from the dataset given and thus observed that the models are predicting the true outputs, except for a few responses.
- Thus I concluded that LLMs are biased slightly and this might be because of the data they are trained on.
- This is previous to proper understanding of the problem given.

- Then I started to work with the dataset given to analyze the bias in LLMs.
- I started to extract information from the .jsonl file using custom parser using "JSON" library and use lists to maintain the true outputs, predicted outputs, identity terms.
- I tried to quantify the bias of LLMs using a simple formula i.e., bias(LLM)=number of biased outputs / total predictions, here for number of biased outputs I have taken the identity terms with the maximum frequency of incorrect predictions of the LLM.
- The order of bias I got from my analysis is -
- Alpha = Gamma > Beta > Delta = Epsilon = Eta = Iota = Theta = Zeta is the decreasing order of bias.