GPU:

Nvidia-sim (check gpu version)

Sanity check Scripts:

python LLM-debiasing.py --train\_data /scratch0/bashyalb/ACL2024/Sanitycheck/Data/shorter\_text.txt --model bert-base-uncased --lr 3e-5

python LLM-debiasing.py --train\_data /scratch0/bashyalb/ACL2024/Sanitycheck/Data/longer\_text.txt --model bert-base-uncased --lr 3e-5

python /scratch0/bashyalb/ACL2024/Sanitycheck/Code/LLM-debiasing.py --train\_data /scratch0/bashyalb/ACL2024/Sanitycheck/Data/longer\_text10k\_gender.txt --model bert-base-uncased --lr 3e-5

jk

nohup python /scratch0/bashyalb/DataAug4SocialBias/Sanitycheck/Code/LLM-debiasing.py --train\_data /scratch0/bashyalb/DataAug4SocialBias/SentenceGeneration/Data/DebiasingCorpus/Orignal/corpus\_1-13\_20k.txt --model bert-base-uncased --lr 3e-5>mylog.log>2&1&

nohup python /scratch0/bashyalb/ACL2024/Sanitycheck/Code/LLM-debiasing.py --train\_data /scratch0/bashyalb/ACL2024/Sanitycheck/Data/longer\_text50k.txt --model bert-base-uncased --lr 3e-5>mylog.log>2&1&

activate conda:

source /anaconda/bin/activate

conda activate software/envc

nohup python mycode.py > mylog.log >2&1 &

ssh -L 2121:bashyalb@voodoo.egr.msu.edu:21 [bashyalb@scully.egr.msu.edu](mailto:bashyalb@scully.egr.msu.edu)

python ../Code/StereoSet.py --debiased\_model debiased.bert.ckpt --mode=debug

**Original(BERT)**

100%|███████████████████████████████████████| 2104/2104 [00:51<00:00, 40.50it/s]

{'lm\_accuracy': 0.8369771863117871, 'st\_accuracy': 0.560361216730038, 'detailed': {'profession': {'lm\_accuracy': 0.8065512978986403, 'st\_accuracy': 0.5550061804697157}, 'race': {'lm\_accuracy': 0.8627858627858628, 'st\_accuracy': 0.5467775467775468}, 'gender': {'lm\_accuracy': 0.8235294117647058, 'st\_accuracy': 0.6274509803921569}, 'religion': {'lm\_accuracy': 0.8782051282051282, 'st\_accuracy': 0.5641025641025641}}}

**10k sentences:**

**Debiased(Short):**

100%|███████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2104/2104 [00:29<00:00, 70.59it/s]

{'lm\_accuracy': 0.7229087452471483, 'st\_accuracy': 0.5147338403041825, 'detailed': {'profession': {'lm\_accuracy': 0.6897404202719407, 'st\_accuracy': 0.5067985166872683}, 'race': {'lm\_accuracy': 0.748960498960499, 'st\_accuracy': 0.5093555093555093}, 'gender': {'lm\_accuracy': 0.7019607843137254, 'st\_accuracy': 0.5411764705882353}, 'religion': {'lm\_accuracy': 0.8141025641025641, 'st\_accuracy': 0.5769230769230769}}}

**Debiased(long):**

100%|███████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2104/2104 [00:37<00:00, 56.02it/s]

{'lm\_accuracy': 0.8001425855513308, 'st\_accuracy': 0.5437262357414449, 'detailed': {'profession': {'lm\_accuracy': 0.7731767614338689, 'st\_accuracy': 0.5278121137206427}, 'race': {'lm\_accuracy': 0.829002079002079, 'st\_accuracy': 0.5457380457380457}, 'gender': {'lm\_accuracy': 0.7686274509803922, 'st\_accuracy': 0.5568627450980392}, 'religion': {'lm\_accuracy': 0.8269230769230769, 'st\_accuracy': 0.6410256410256411}}}

50k sentences:   
  
Debiased (Short):

100%|███████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 2104/2104 [00:29<00:00, 72.22it/s]

{'lm\_accuracy': 0.7112642585551331, 'st\_accuracy': 0.5261406844106464, 'detailed': {'profession': {'lm\_accuracy': 0.6687268232385661, 'st\_accuracy': 0.5203955500618047}, 'race': {'lm\_accuracy': 0.7458419958419958, 'st\_accuracy': 0.5166320166320166}, 'gender': {'lm\_accuracy': 0.696078431372549, 'st\_accuracy': 0.5450980392156862}, 'religion': {'lm\_accuracy': 0.7756410256410257, 'st\_accuracy': 0.6410256410256411}}}

Debiased (Long):

{'lm\_accuracy': 0.7825570342205324, 'st\_accuracy': 0.5418250950570342, 'detailed': {'profession': {'lm\_accuracy': 0.7552533992583437, 'st\_accuracy': 0.5327564894932015}, 'race': {'lm\_accuracy': 0.8108108108108109, 'st\_accuracy': 0.5436590436590436}, 'gender': {'lm\_accuracy': 0.7509803921568627, 'st\_accuracy': 0.5411764705882353}, 'religion': {'lm\_accuracy': 0.8205128205128205, 'st\_accuracy': 0.6153846153846154}}}

* 3 kind of sentences in the corpus:
  + Shorter
  + Longer with social group and associated attributes.
  + Longer with no social group and associated attributes
* Identifying two or more social groups in the corpus:
  + Select some examples containing two social groups for in context learning training.
  + Design prompt accordingly,
  + Query each sentence with LLMs
* Once the sentences with the social groups are identified:
  + CDA with LLMs based on the social group
  + This can be done at the same time during previous experiment or later separately
* Similarly, perform the other augmentation:
  + Short sentences with no context
  + If required, other long sentences with no social groups
* Experiments:
  + SG CDA
  + SG CDA + other long sentence
  + SG CDA+ all sentences
* SG CDA+ other long sentences+ aug sentences (augmented short sentences)

Reasoning and Explanation:

Reasoning for Social group classification:

Identify the sentences with social groups

* Brute force

Sentence generation

Explanation for enhancing diversity in sentences with social groups:

Todo (Nov 12, Sunday)

Measure the bias in the generated sentences, design some metrics

Checking the sentence generation with the sentences with different social groups:

* Gender (Test on the 100 quality)
* Race (same)
* Quality and diversity metrics
* Compare no. of tokens in the corpus

Also, there will be 9 extra sentences

Do we need to augment the other sentences which are long without any social group.

To run AttenD: use the environment new\_env  
  
  
  
  
  
  
Measure the original corpus with the corpus with same social group.   
Metrics for measuring the quality of gender representation in CDA:   
1. Word frequency analysis on the both the gender corpus

Generate 4 more sentences using both corpus:

What is metrics for measuring diversity and metrics of quality.

Debiasing Results:

Corpus: 10-40, 10 k

BERT:

{'lm\_accuracy': 0.8369771863117871, 'st\_accuracy': 0.560361216730038, 'detailed': {'profession': {'lm\_accuracy': 0.8065512978986403, 'st\_accuracy': 0.5550061804697157}, 'race': {'lm\_accuracy': 0.8627858627858628, 'st\_accuracy': 0.5467775467775468}, 'gender': {'lm\_accuracy': 0.8235294117647058, 'st\_accuracy': 0.6274509803921569}, 'religion': {'lm\_accuracy': 0.8782051282051282, 'st\_accuracy': 0.5641025641025641}}}

Original corpus(50 epochs):   
{'lm\_accuracy': 0.7604562737642585, 'st\_accuracy': 0.5185361216730038, 'detailed': {'profession': {'lm\_accuracy': 0.7441285537700866, 'st\_accuracy': 0.5105067985166872}, 'race': {'lm\_accuracy': 0.7744282744282744, 'st\_accuracy': 0.498960498960499}, 'gender': {'lm\_accuracy': 0.7509803921568627, 'st\_accuracy': 0.5882352941176471}, 'religion': {'lm\_accuracy': 0.7884615384615384, 'st\_accuracy': 0.6153846153846154}}}

Generated corpus (50 epochs):

{'lm\_accuracy': 0.8025190114068441, 'st\_accuracy': 0.5228136882129277, 'detailed': {'profession': {'lm\_accuracy': 0.7669962917181706, 'st\_accuracy': 0.5203955500618047}, 'race': {'lm\_accuracy': 0.8321205821205822, 'st\_accuracy': 0.498960498960499}, 'gender': {'lm\_accuracy': 0.7901960784313725, 'st\_accuracy': 0.596078431372549}, 'religion': {'lm\_accuracy': 0.8461538461538461, 'st\_accuracy': 0.6025641025641025}}}

Original CDA corpus (50 epochs):

{'lm\_accuracy': 0.8486216730038023, 'st\_accuracy': 0.5518060836501901, 'detailed': {'profession': {'lm\_accuracy': 0.8207663782447466, 'st\_accuracy': 0.5611866501854141}, 'race': {'lm\_accuracy': 0.8716216216216216, 'st\_accuracy': 0.5228690228690228}, 'gender': {'lm\_accuracy': 0.8431372549019608, 'st\_accuracy': 0.6274509803921569}, 'religion': {'lm\_accuracy': 0.8717948717948718, 'st\_accuracy': 0.5641025641025641}}}

Generated CDA corpus (50 epochs) :

{'lm\_accuracy': 0.7647338403041825, 'st\_accuracy': 0.5076045627376425, 'detailed': {'profession': {'lm\_accuracy': 0.745982694684796, 'st\_accuracy': 0.511742892459827}, 'race': {'lm\_accuracy': 0.7827442827442828, 'st\_accuracy': 0.4802494802494803}, 'gender': {'lm\_accuracy': 0.7411764705882353, 'st\_accuracy': 0.5764705882352941}, 'religion': {'lm\_accuracy': 0.8141025641025641, 'st\_accuracy': 0.5769230769230769}}}