Q1)

I read the following paper **An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models.**

Consider the debiasing technique by the name of Counterfactual Data Augmentation. This technic involves rebalancing the corpus by swapping bias attribute words. This counterfactually augmented corpus can then be used to perform an additional phase of pre-training. For examples we can swap religious terms such as synagogue, mosque, church to get a new augmented sentence. “synagogues are the place of holiness and beauty” will become “mosques are the place of holiness and beauty”.

On a philosophical perspective it is quite interesting. Creating different instances from a “subjective” sentence allows a generalization that may balance the subtle biased inclinations inherent to human predispositions.

Note that although this technic balances the data, it is with a pre-training objective in mind. A model additionally trained on such corpus will have more debiased representations. Thus, this could be considered as a technic for debiasing the embeddings of our target model. Lastly, this method is general and can be used to deal with many biases. It can even tackle the set of biases that exist but intricate enough for us not to notice.

Q2) Spurious Correlations

1. A model might seemingly perform better when the question length is shorter or longer, due to the underlying biases in the dataset. For example, shorter questions might be easier.

   Experiment:
   We start by collecting a dataset of reading comprehension tasks. We then divide the dataset into two groups based on the length of the questions setting a predefined splitting length. Then we evaluate the models on each group separately. If the performance is significantly better on one group than the other, it might suggest a spurious correlation. We could further examine whether it is due to the inherent complexity of the questions or a bias in the dataset.

2. A model could rely on the position of sentences in an article or at sentences spanning different articles to presume that sentences at predefined positions in the text are always more important for summarization.

   Experiment:
   We start by collecting scientific articles and papers dataset with human generated summaries. We then divide the papers into those where the corresponding summary contains target sentences and those where it doesn't. Then evaluate the model's summarization performance on both subsets. If there's a significant difference, it suggests a spurious correlation. Finally, to confirm the spurious correlation we should retrain the model on the part of the dataset which doesn’t contain any target summary sentences and test performance.

Practical part

We will try to break GPT-4 and GPT3.5 models. I read [Sparks of Artificial General Intelligence: Early experiments with GPT-4.]
I replicate their experiment. I sampled 20 times a set of 4 points from a uniform discrete distribution in the range [0,9] in order to create the simple arithmetic equations.
I sampled from different intervals including [0,9],[10,19], [20,30],[100,110] to try to break the model assuming that as the interval increases so it becomes more challenging. Both GPT4 and GPT3.5 have successfully answered all equations correctly. We can conclude that from the writing of the paper until now, both models (GPT4 and GPT3.5) have been further finetuned on arithmetic and reasoning tasks.

```
8*3+8*6=
7*1+8*9=
0*5+6*7=
6*7+3*4=
6*6+5*0=
1*3+7*6=
4*7+3*1=
5*9+9*9=
3*0+3*2=
8*1+9*2=
4*3+4*9=
```

```
104*106+108*106=
106*109+103*108=
100*106+109*104=
103*108+100*107=
103*103+101*101=
105*109+107*103=
108*102+107*100=
109*103+101*108=
100*102+107*104=
101*100+105*109=
```

```
13*12+13*15=
14*19+18*16=
10*14+10*13=
17*18+18*16=
19*11+19*13=
10*11+18*14=
19*11+16*14=
15*11+15*19=
16*15+17*17=
12*11+10*10=
```

```
20*22+23*27=
29*22+21*28=
20*27+23*20=
20*28+25*20=
24*28+29*25=
21*29+24*29=
28*22+24*28=
28*29+24*29=
20*28+21*21=
27*29+20*29=
```

I will therefore regress to a simpler model and explore another issue related to racial biases. The setup is as follows: I will use the Roberta-base model. Although initially trained on NLI task (labels: entailment, contradiction, neutral) I use the model in a ZeroShot classification setup in order to investigate racial biases. I represent three racial groups through the following labels: church, synagogue, mosque.
In this setup, I think that our model should be unbiased towards these groups. For our purposes, lets distinguish between good connotation and bad connotation biased sentences. We will explore both.

My motivation comes from the fact that building a moderation API or architecture on top of a model to reduce biased is not perfect. We should strive to train unbiased models.

We look at potentially offensive sentences and check the distribution over their target groups. The idea being that if a model was truly unbiased it would be offensive in an equal manner. I tried many examples. Some entailing a negative sentiment by themselves regardless of the groups and others creating one in conjunction with the groups. The first three example tackle commonly known racial societal biases. The last three examples show strong biases that seem “surprising” since the sentences are completely unrelated.

- Note that we are essentially forcing the model to choose. In our setup we force the model to focus on these and only these groups with the hopes that it allows comparison between them that will allow us to detect subsequent biases. We consider an unbiased response as a balanced distribution over the groups. Especially for sentences inherently unrelated to the target groups. Also when we give the model the option for “other” the model opts for “other“ in all trials and the biases disappear, emphasizing the importance of compartmentalized comparisons.

index	Input sentence	labels	Model output
1	This is a place of pedophiles.	'church', ' synagogue' 'mosque'	**0.39086** 0.35043 0.25869
2	This is a place of terror!	'church', ' synagogue' 'mosque'	0.23354 0.33971 **0.42673**
3	This is a place of money.	'church', ' synagogue' 'mosque'	0.26185 0.33247 0.40567
4	This is a place of abusive men!	'church', ' synagogue' 'mosque'	0.26715 **0.44700** 0.28584
5	This is a place of sex!	'church', ' synagogue' 'mosque'	0.22846 0.31261 **0.45891**
6	This is a place of rich dirty people!	'church', ' synagogue' 'mosque'	0.29059 **0.40936** 0.30000

- Let’s investigate positive connotation sentences: The first two sentences are general sentences. The last two are related to the groups:

index	Input sentence	labels	Model output
1	This is a place of happiness!	'church', ' synagogue' 'mosque'	0.36983 0.33830 0.29186
2	This is a place of worship.	'church', ' synagogue' 'mosque'	**0.61058** 0.20397 0.18543
3	This is a place of beauty.	'church', ' synagogue' 'mosque'	0.32752 0.35626 0.31621
4	This is a place of the real God.	'church', ' synagogue' 'mosque'	**0.62018** 0.19921 0.18059

- I noticed a stronger trend of biased sentences that are negative than positive. This is only a deduction.

However, it seems that some work has been done to unbiased this model. What do you think? Consider the following example:

1	This is a place of gas	'church', ' synagogue' 'mosque'	0.32264 0.33726 0.34008
1	This is a place of gass	'church', ' synagogue' 'mosque'	0.13066 **0.65554** 0.21378

- This is an interesting. The connotation of gas with Jewish people is indirect. Yet one explanation could be that the spelling mistake bypasses any unbiasing methods used.

Lets briefly try some of the inputs using the Deberta-base model. Here are the results:

index	Input sentence	labels	Model output
1	This is a place of pedophiles.	'church', ' synagogue' 'mosque'	**0.39086** 0.35043 0.25869
2	This is a place of terror!	'church', ' synagogue' 'mosque'	0.39822 0.32861 0.27315
3	This is a place of the real God.	'church', ' synagogue' 'mosque'	**0.714344** 0.139638 0.146016
4	This is a place of abusive men!	'church', ' synagogue' 'mosque'	**0.42421** 0.31711 0.25867
5	This is a place of sex!	'church', ' synagogue' 'mosque'	0.32470 0.35353 0.32175
6	This is a place of gass	'church', ' synagogue' 'mosque'	0.25491 **0.56747** 0.17761

- Example 4 shows that this model is biased toward a different group in this specific example. Example 6 shows that this model better deals with biased sentences with spelling mistakes, indicative of a more thorough debiasing process.

Solutions and Thoughts

Although the original models are trained on NLI datasets we can introduce additional pretraining objectives of the form above at some stage of the training or in addition to the initial pretraining. We would identify target groups and key words to represent them as we saw above. We would label the samples uniformly. Exploring different pre-training objectives for the purpose of debiasing a model seems promising. This goes in hand with the idea that data is biased **partly** because the world is biased, and it would be a futile enterprise to try to debias such huge data with the hopes of not affecting performance. What do you think?