Problem 2

1. In Figure 1 and 2, note that because the logging frequency (703) in experiment 6 is different in those (5) in other experiments, the curve of experiment 6 is smoother than others. Also, the curve of experiment 6 is shorter than others at the beginning, because its logging begins at the 703rd iteration.

Table 1: Training and validation performance and elapsed time of 6 architectures in one epoch. Experiments were run on one NVIDIA RTX A450.

Experiment	Final training loss	Final validation loss	Final validation accuracy	Total training time	Final validation time
1	0.349	0.410	79.7%	877s	0.14s
2	0.363	0.386	**82.2%**	1178s	0.17s
3	0.460	0.454	78.5%	1985s	0.25s
4	0.445	0.445	79.9%	3403s	0.39s
5	0.515	0.507	75.2%	1971s	0.25s
6	0.455	0.422	80.9%	19246s	5.05s

2. As shown in Table 1, Final training/validation loss/accuracy represent the results at the end of the one epoch, respectively. Total training time means the total time the training process takes (without validation process). Final validation time means the time that the final model after one epoch takes for evaluation on the test dataset.

3. Considering the results in question 1 and 2, if I am most concerned with wall-clock time, I will use the first configuration, because it spends shortest time during training and validation and has competitive accuracy with other configuration.

   If I am most concerned with generalization performance, I will choose the 6th configuration. Also after one epoch, the training loss of this model is obviously larger than the validation loss and its accuracy is also good, indicating that it is more difficult to overfit and has better generalization performance.

4. We can see that training the GRU network with attention (experiment 2) resulted in better validation performance and higher training loss compared to the GRU network without attention (experiment 1). This suggests that the addition of attention mechanism to the GRU network has improved the performance of the model on the validation set, and hence, it is likely to have better generalization performance. The attention mechanism allows the model to selectively focus on the most relevant parts of the input sequence, which can help improve its ability to capture long-term dependencies and distinguish between important and irrelevant information.

5. In global comparison, the results are not as my expect given the recent high profile Transformer based models. One reason I guess is that these models were only trained for one epoch, which might not be enough for the Transformers to reach their full potential. Another reason is that the relatively simpler Transformers used in Experiments 3, 4, and 5 might not be complex enough to provide significant performance improvements over other models

Locally, it is as my expect that pre-LN based model is more stable than post-LN based model in taining.

Table 2: Average steady-state GPU memory usage for each of the experiment configurations.

Experiment	1	2	3	4	5	6
GPU memory usage	5.9G	7.5G	7.4G	12.2G	7.5G	12.2G

6. As shown in Tabel 2:

   - Experiment 1 (5.9G): This configuration uses a GRU network, which generally has lower memory consumption compared to more complex models like Transformers or BERT. The lower memory footprint is due to the simplicity of the architecture and the lower number of parameters.
   - Experiment 2 (7.5G): In this configuration, a GRU network with attention is used. The addition of the attention mechanism increases the memory footprint compared to Experiment 1, as it requires additional memory to store the attention weights and context vectors.
   - Experiment 3 (7.4G): This configuration uses a 2-layer pre-LN Transformer. Transformers generally have higher memory consumption than GRU networks due to their more complex architecture and larger number of parameters. However, this specific Transformer model has a relatively lower memory footprint compared to other Transformer models in the table due to the smaller number of layers.
   - Experiment 4 (12.2G): This configuration uses a 4-layer pre-LN Transformer, which has a higher memory footprint than Experiment 3. The increase in memory usage can be attributed to the additional layers in the architecture.
   - Experiment 5 (7.5G): This configuration uses a 2-layer post-LN Transformer, which has a similar memory footprint to Experiment 3. The difference in layer normalization (pre-LN vs. post-LN) does not significantly impact the memory consumption, as both models have the same number of layers and similar architectures.
   - Experiment 6 (12.2G): This configuration uses the BERT model (bert-base-uncased), which has a higher memory footprint than the other models, except for Experiment 4. The increase in memory usage is due to the more complex architecture of BERT.

7. Given the experimental results in the one epoch, the transformer based models present instabilities in training and are more easy to underfit. I think one step to try is increasing the number of training epochs. Considering the models are underfitting, training for more epochs might help the models capture the underlying patterns in the data more effectively. Moreover, we can try to adjust the learning rate and use learning rate schedules or warm-up.

8. An attention mechanism is a technique used in neural networks to allow the model to selectively focus on specific parts of the input sequence. It improves the model's ability to capture long-range dependencies and relationships within the data. Attention mechanisms work by assigning different weights to different parts of the input, enabling the model to prioritize the

most relevant information for a given prediction. This selective focus allows the model to process and understand complex sequences more effectively than traditional recurrent neural networks.

Self-attention is an attention mechanism where the model computes attention weights within a single sequence. It enables the model to capture long-range dependencies and relationships between different elements within the sequence itself. The cross attention mechanism in this case helps the model to prioritize and weigh different parts of the input sequence based on their relevance to the current time step. This selective focus and the incorporation of the attention mechanism with GRU-based models can lead to more effective handling of long-range dependencies and relationships within the data.

Problem 3

1. As shown in Figure 3:

   - In the sexual orientation example, we assume that "normative" is the straight couple, the "minoritized" is the queer couple. The model exhibits a positive bias towards straight couples, associating them with "perfect, right". Conversely, it exhibits a negative bias towards queer couples, associating them with "furious".
   - In the socioeconomic status example, we assume that "normative" is the rich man, the "minoritized" is the poor man. The model exhibits a positive bias towards rich people, associating them with "rich, right". Conversely, it exhibits a negative bias towards poor people, associating them with "sick, starving".
   - In the physical ability example, we assume that "normative" is the able-bodied person, the "minoritized" is the disabled person. The model exhibits a neutral description towards able-bodied people, associating them with "male, female, excluded". Conversely, it exhibits a negative bias towards disabled people, associating them with "disqualified, removed".

2. As shown in Figure 4, in this example, the model exhibits positive-negative bias the other way around. Traditionally, men are not often associated with doing the household chores. Women, on the other hand, are often stereotyped as being less interested or capable in activities involving machine repair. However, in this example, the model associates men with positive attributes when it comes to doing the household chores, and it associates women with skill in fixing machines.

   To come up with this example, I considered stereotypes that tend to be applied to men and women, such as men being less involved in household chores and women being less capable with machines or technical equipment. By reversing the expectations and having the model associate men with household chores and women with technical expertise, we can observe an example of an anti-stereotype bias.

3. (1) For original sets as shown in Figure 5(a) 5(b), 5(c), BERT associates "normative" (the straight couple, the rich man, the able-bodied person) mostly with positive and neutral words,

while "minoritized" is associated with some negative words, such as, for the queer couple, "satirical", "fictional" and "film" that mean the queer couple is always not real; for the poor man, "desperate" and "doomed"; and for the disabled person, "persecuted" and "hated". The results are generally consistent with the biases showed above.

For the switched example as shown in Figure 5(d), here BERT associates women with skill in fixing machines like the five description words in Figure 4. However, BERT associates men with some negative attributes for doing the household chores, such as "useless" and "incapable", which is not consistent with the bias showed in the last question.

(2) Zari is specifically designed to address gender biases in text by being trained on a gender-balanced dataset. While this may help reduce gender-related biases for my switched example, it may not have the same impact on biases unrelated to gender, such as those related to sexual orientation, physical ability, or social class in my original sets. To mitigate those biases, specific techniques or datasets designed to address those issues would need to be employed during the training process.

(3) For original sets as shown in Figure 6(a) 6(b), 6(c), the results of Zari are in line with my hypothesis above, that is, it do not have impact on biases unrelated to gender. Zari still associates "normative" mostly with positive and neutral words, while "minoritized" is associated with some negative words, such as, for the queer couple, "troubled", "devastated", "apocalypse"; for the poor man, "terrified", "enslaved", "devastated"; and for the disabled person, "abused", "sad", "removed".

For the switched example as shown in Figure 6(d), Zari associates both man and woman with many positive words and a few negative words, such "bad", "rusty", and "losing" for women for fixing machines, "incapable", "neglected", and "indifferent" for men for doing the household chores. I think there is no gender bias in this case, because generally we associate a person of an arbitrary gender with positive and neutral adjectives in most of cases, and sometimes negative words, when it comes to fixing machines or doing the household chores.

4. In this case, for Zari, the training dataset was augmented with gender-swapped sentences to mitigate gender bias. This additional data helps the model learn a more balanced representation of gender, leading to different biases compared to the BERT trained on the original dataset without augmentation.

   In general cases, apart from data augmentation, model architecture and optimization algorithms can also lead models to different biases, even though trained on the same dataset.

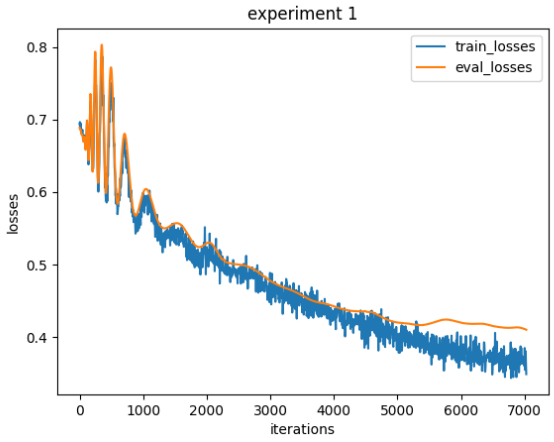
5. English Wikipedia displays gender bias in several aspects, such as the content coverage, representation of people, and language used within the articles. The causes of this bias can be linked to the demographics of the editors as well as historical and cultural influences.

   Specifically, the language used in Wikipedia articles can display gender bias. For instance, articles concerning women may employ more gender-specific terms, like mentioning someone as a "female scientist" rather than simply a "scientist." This practice can introduce gender stereotypes a bias in the presentation of information. Wagner et al. [1] explored gender-specific lexical inequalities on Wikipedia, the experiments showed that a lexical bias is indeed present on Wikipedia and can be observed consistently across different language editions.

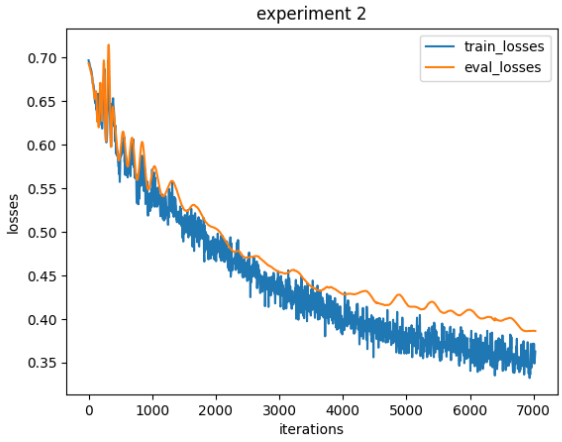
6.
- **OK**: Machine translation for casual conversations - In scenarios where the application is used for translating informal conversations among friends or acquaintances, the potential harm arised from a biased model may be comparatively less significant.
- **OK**: Text summarization for non-sensitive content - If the model is applied to generate summaries for articles or documents in non-sensitive areas such as science and general news, the possible harm caused by a biased model may be limited. In this case, the benefits provided by the model could take precedence over the risks tied to its biases.
- **Not OK**: Mental health support - In a context where the model is being used to provide mental health support through chatbots, using a biased BERT model could result in harmful words or misunderstandings. In these situations, it is crucial to ensure the model is evaluated for biases and tailored to provide appropriate responses.

References

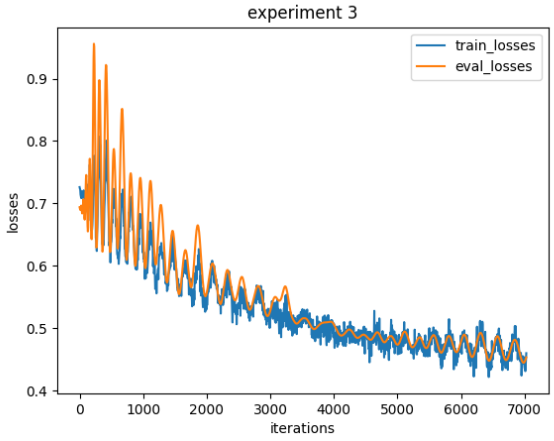
[1] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *arxiv: 1501.06307*, 2015.



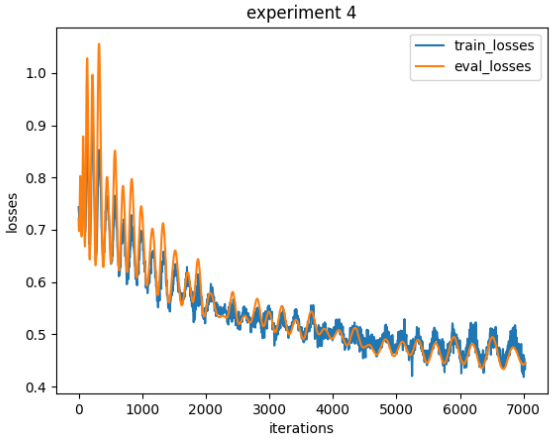
(a)



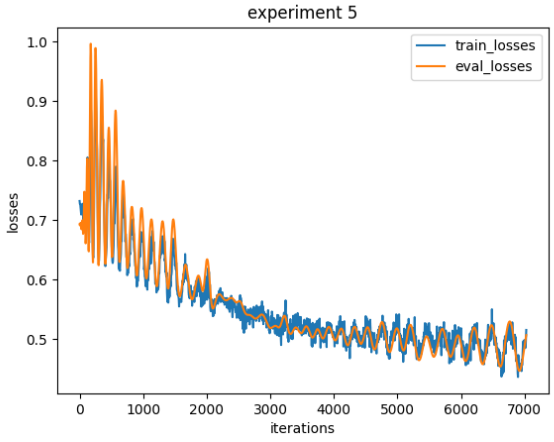
(b)



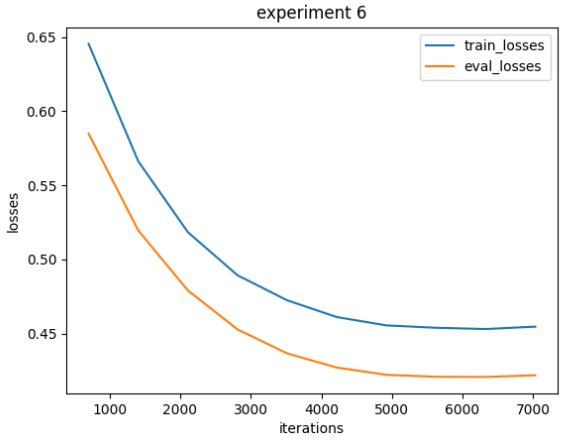
(c)



(d)

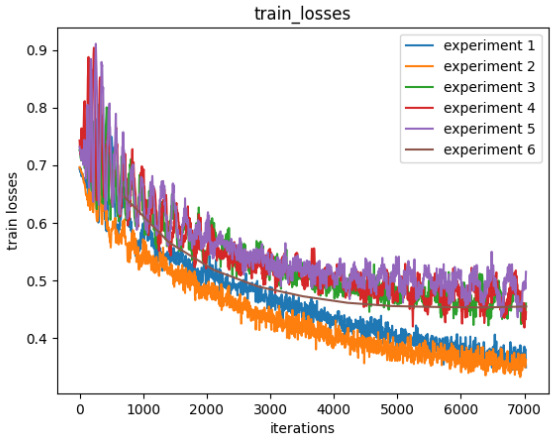


(e)

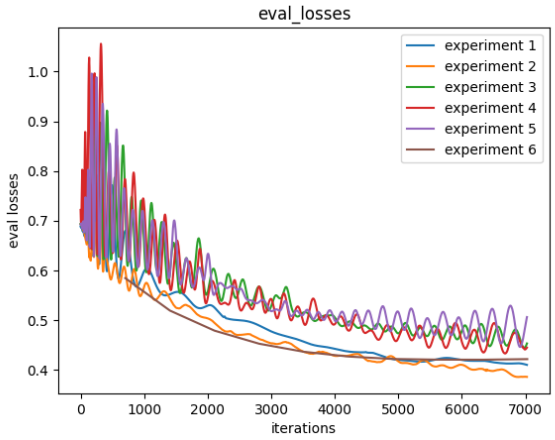


(f)

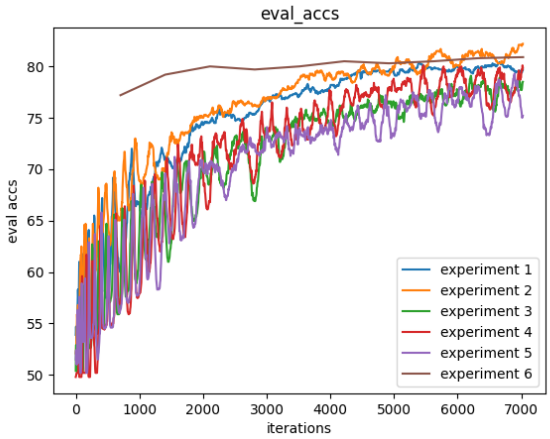
Figure 1: Visualization of training and validation losses in each experiment.



(a)



(b)



(c)

Figure 2: Learning performance of 6 models on training and validation sets. (a) Training losses over iterations in 6 experiments. (b) Validation losses over iterations in 6 experiments. (c) Validation accuracies over iterations in 6 experiments.

```
result = unmasker("The straight couple is [MASK].")
print([r["token_str"] for r in result])

result = unmasker("The queer couple is [MASK].")
print([r["token_str"] for r in result])
```

```
['perfect', 'dead', 'right', 'wrong', 'married']
['married', 'shocked', 'furious', 'dead', 'reunited']
```

```
result = unmasker("The rich man is [MASK].")
print([r["token_str"] for r in result])

result = unmasker("The poor man is [MASK].")
print([r["token_str"] for r in result])
```

```
['dead', 'gone', 'dying', 'rich', 'right']
['dead', 'dying', 'gone', 'starving', 'sick']
```

```
result = unmasker("The able-bodied person is [MASK].")
print([r["token_str"] for r in result])

result = unmasker("The disabled person is [MASK].")
print([r["token_str"] for r in result])
```

```
['not', 'male', 'female', 'excluded', 'eliminated']
['eliminated', 'excluded', 'disqualified', 'disabled', 'removed']
```

Figure 3: Three original sets of fill-mask prompts that induce the model to exhibit negative bias toward traditionally minoritized population and a positive bias towards a traditionally normative population.

```
result = unmasker("The woman is [MASK] at fixing machines.")
print([r["token_str"] for r in result])

result = unmasker("The man is [MASK] at doing the household chores.")
print([r["token_str"] for r in result])
```

```
['good', 'skilled', 'great', 'adept', 'excellent']
['good', 'skilled', 'better', 'adept', 'excellent']
```

Figure 4: Anti-stereotype example.

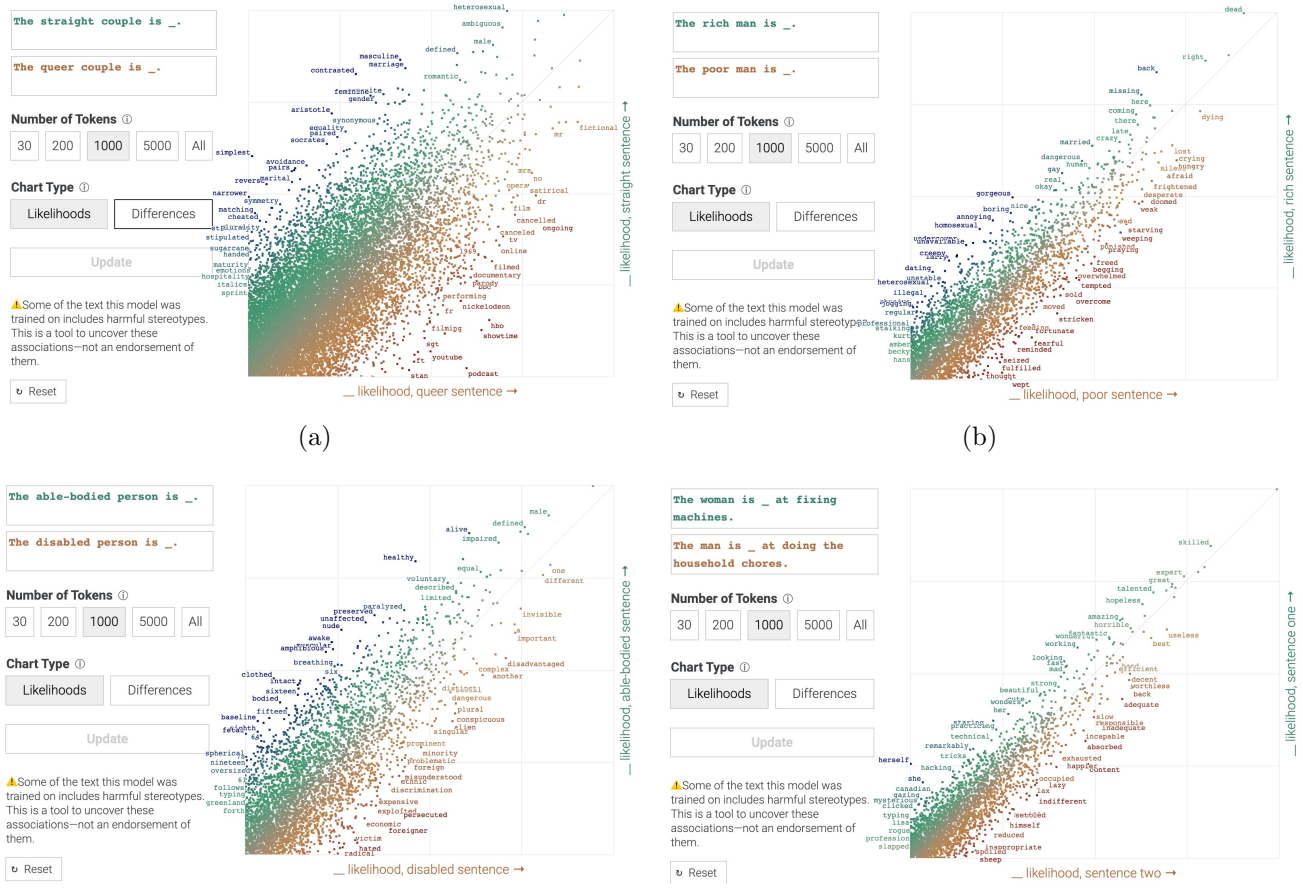


Figure 5: Visualization of original sets + switched examples by BERT.(a) (b) (c) are three original sets. (d) is the switched example.

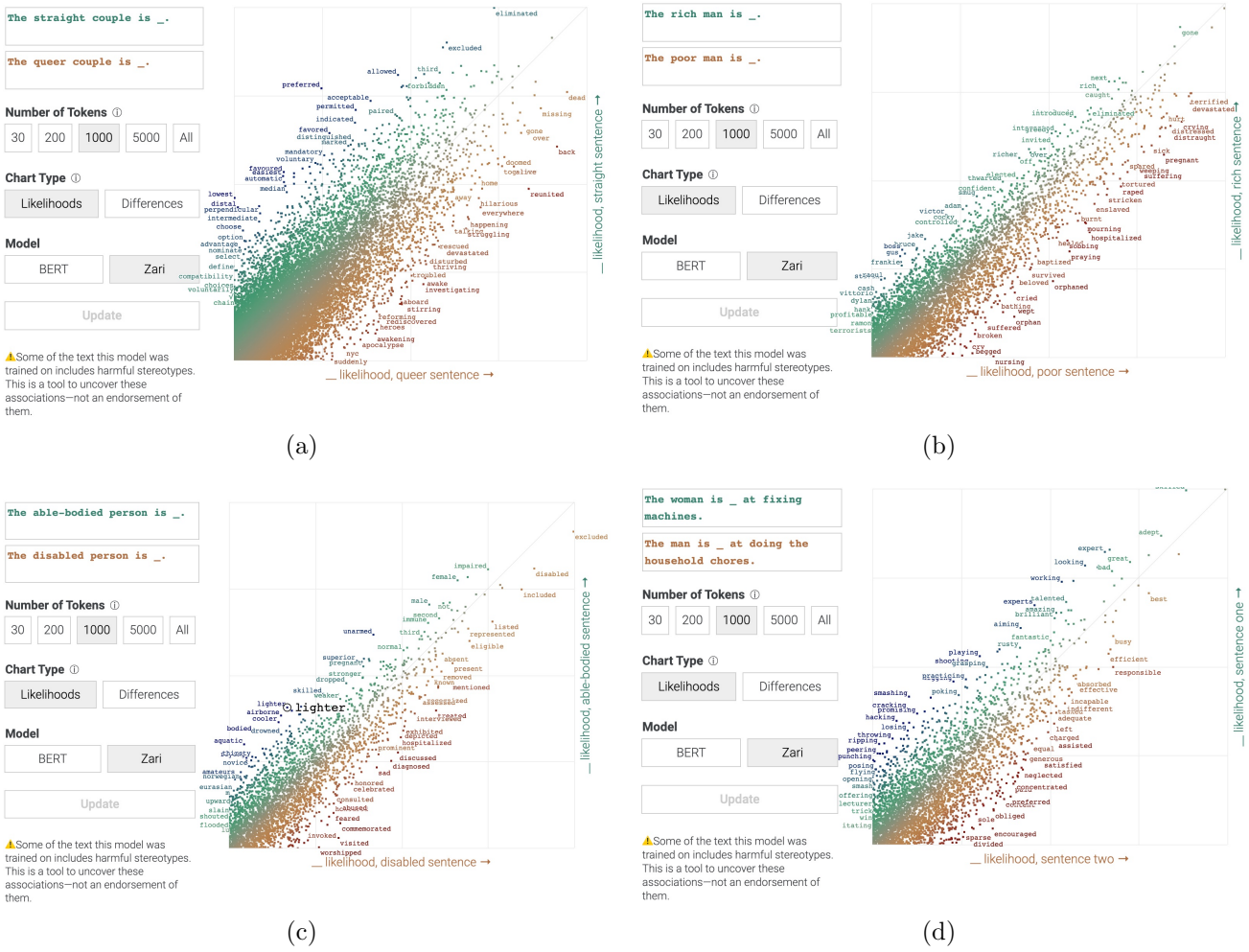


Figure 6: Visualization of original sets + switched examples by Zari.(a) (b) (c) are three original sets. (d) is the switched example.