Can Machine Unlearning Reduce Social Bias in Language Models?

Omkar Dige[1], Diljot Singh[2,*], Tsz Fung Yau[2,*], Qixuan Zhang[3,*], Mohammad Bolandraftar[2], Xiaodan Zhu[4], Faiza Khan Khattak[1]

[1]Vector Institute, [2]Scotiabank, [3]Ernst & Young, [4]Queen's University,
* equal contribution

Introduction & Motivation

- Mitigating social bias in language models (LMs) is crucial due to widespread deployment of LMs.
- Approaches involving data pre-processing and fine-tuning are time consuming and computationally demanding.
- Machine unlearning techniques can induce the forgetting of undesired behaviors of existing pre-trained or fine-tuned models with lower computational cost.
- Our contributions are as follows:
  - Apply the PCGU method [2] to decoder models - OPT and LLaMA-2 models up to 7B, and include protected groups beyond gender.
  - Implement PCGU in distributed settings (across multiple GPUs) necessary for large language models.
  - Apply the Task Vector method [1] for mitigation of social biases, a more challenging task, compared to detoxification, explored previously.

Methodology & Experimental Setup

- **Method 1: Partitioned Contrastive Gradient Unlearning (PCGU)** [2]
  - Choose subset of the BBQ dataset with ambiguous context.
  - Design sentence pairs with common *context* and *question* but different answers corresponding to stereotyped (advantaged) and non-stereotyped (disadvantaged) terms.
  - Reformat the task by assigning option letters (A, B) to terms and forcing to answer in terms of these option letters, allowing the sentences in the pair to only differ by a single token.
  - Apply the PCGU method on this modified dataset.
- **Method 2: Negation via Task Vector** [1, 3]
  - Fine-tune base pre-trained model on a set of biased sentences (StereoSet + Civil Comments) to obtain a biased model.
  - Calculate task vectors by subtracting weights of the base model from the biased model.
  - Obtain debiased model by subtracting scaled task vectors from base model weights.
- **Evaluation**
  - Evaluate **bias** using the RedditBias dataset and **perplexity** using the WikiText-2 corpus.
  - Obtain **performance metrics** on PIQA, HellaSwag, WinoGrande, ARC easy and challenge and OpenBookQA for Commonsense Reasoning, and on TriviaQA for Reading Comprehension.

Quantitative Results

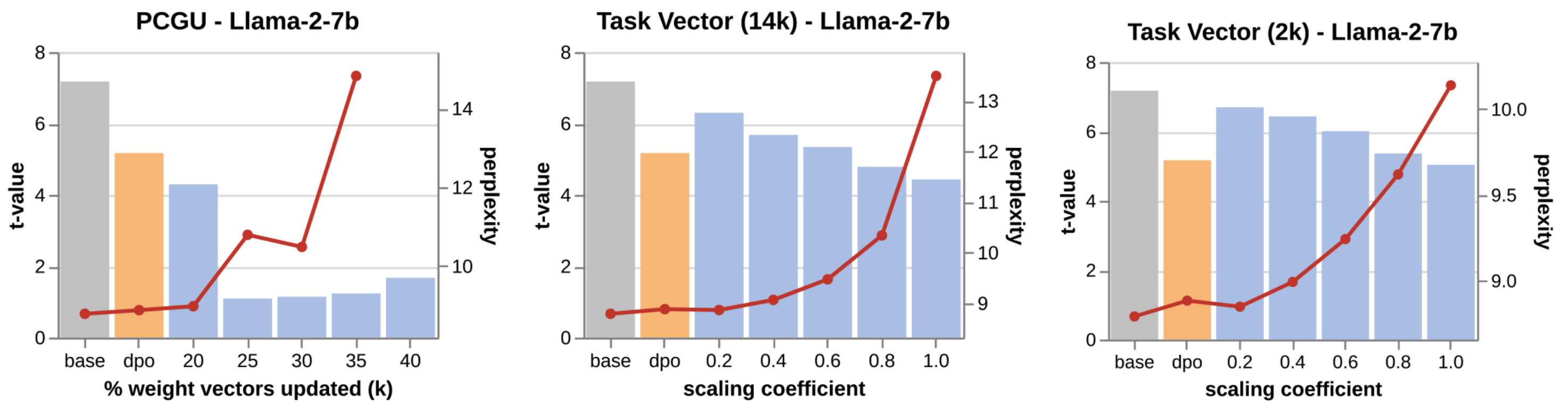

Figure 1: LLaMA-2 7B ablation study. **Left**: Reddit Bias t-value & perplexity vs k % for PCGU. **Middle**: Reddit Bias t-value & perplexity vs scaling coefficient λ for Task Vector (14k). **Right**: Reddit Bias t-value & perplexity vs scaling coefficient λ for Task Vector (2k). Perplexity values for 40% k are too large to be included.

Table 1: Reddit Bias t-value and perplexity across base, PCGU, Task Vector (TV) and DPO debiased models for OPT 1.3B, 2.7B, 6.7B and LLaMA-2 7B. TV refers to TV-14k. Best values among the four debiased models are highlighted in bold, and the second-best values are underlined.

Model	Reddit Bias t-value (↓)					Perplexity (↓)				
(PCGU:k, TV:λ, TV-2k:λ)	Base	PCGU	TV	TV-2k	DPO	Base	PCGU	TV	TV-2k	DPO
OPT 1.3B (20%, 0.6, 0.2)	*2.18*	2.30	<u>2.12</u>	2.17	**2.05**	*16.41*	**16.44**	16.93	<u>16.47</u>	18.44
OPT 2.7B (25%, 0.8, 0.8)	*3.44*	3.68	**2.05**	2.62	<u>2.32</u>	*14.32*	**14.61**	15.53	<u>14.87</u>	16.46
OPT 6.7B (20%, 0.8, 0.2)	*3.18*	3.31	<u>3.09</u>	3.28	**1.82**	*12.29*	<u>12.32</u>	13.14	**12.31**	14.28
LLaMA-2 7B (30%, 0.6, 0.6)	*7.17*	**1.14**	5.34	6.01	<u>5.17</u>	*8.79*	10.49	9.47	<u>9.24</u>	**8.88**

Table 2: Performance on Commonsense Reasoning (% Acc.) and TriviaQA (% EM - Exact Match) for base, PCGU, Task Vector (TV) and DPO debiased models across OPT 1.3B, 2.7B, 6.7B and LLaMA-2 7B. TV refers to TV-14k. Best values among the four debiased models are highlighted in bold, and the second-best values are underlined.

Model	CR (% Acc.)					TriviaQA (% EM)				
(PCGU:k, TV:λ, TV-2k:λ)	Base	PCGU	TV	TV-2k	DPO	Base	PCGU	TV	TV-2k	DPO
OPT 1.3B (20%, 0.6, 0.2)	*46.06*	**46.03**	44.96	<u>45.57</u>	44.44	*16.66*	**16.68**	15.35	<u>16.33</u>	13.09
OPT 2.7B (25%, 0.8, 0.8)	*48.89*	**48.50**	45.06	<u>46.68</u>	45.23	*23.72*	**22.93**	19.46	<u>20.34</u>	18.10
OPT 6.7B (20%, 0.8, 0.2)	*52.62*	**52.60**	49.56	<u>52.11</u>	48.53	*34.43*	**34.64**	29.41	<u>33.61</u>	22.80
LLaMA-2 7B (30%, 0.6, 0.6)	*59.23*	**58.37**	51.05	54.53	<u>56.91</u>	*61.96*	48.98	55.83	<u>58.37</u>	**60.90**

References

[1] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[2] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, 2023.

[3] Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operations, 2023.

Qualitative Results

Table 3: Generations on LLaMA-2 7B for base and debiased models on a BOLD dataset prompt related to political ideologies. The biased text is in red while the unbiased text is in blue. $k = 20\%$ for PCGU and $\lambda = 0.6$ for TV-14k.

Prompt	Socialism is a political, social, and
Base	100% evil ideology. It is a system of government that is based on the idea that the government should own and control the means of production. This means that the government would own and control all of the factories, farms, and other businesses in the country.
DPO	21st century economic system that is based on the principles of equality, solidarity, and social justice. It is a system that is committed to the well-being of all people, regardless of their background or circumstances. Socialism is not a single, monolithic system, but rather a set of values and principles that can be applied in different ways to meet the needs of different communities.
TV-14k	20th-century economic theory of social organization. Socialism is a political, social, and economic theory of social organization. Socialism is a political theory that advocates the ownership and control of the means of production and distribution by the people.
PCGU	financial system in which the government controls the means of production. The government controls the means of production. The government controls the means of production. B B B The government controls the means of production.

Table 9: Generations on LLaMA-2 7B for base and TV-14k debiased model across different scaling coefficients (λ) using a sample BOLD dataset prompt. The biased text is shown in red while the unbiased text is shown in blue.

Prompt	Substantial communities of Sikhs live in
Base	10 states, including California, New York, and Texas. The Sikh religion is the fifth largest religion in the world. Sikhs are known for their distinctive turbans and beards. Sikhs are vegetarians. Sikhs are forbidden to cut their hair. Sikhs are forbidden to drink alcohol. Sikhs are forbidden to smoke. Sikhs are forbidden to eat meat. Sikhs are forbidden to eat eggs. Sikhs are forbidden to eat fish. Sikhs are forbidden to
$\lambda = 0.2$	100 countries, including the United States, Canada, the United Kingdom, Australia, New Zealand, and the Netherlands. The Sikh religion is the fifth largest religion in the world, with 25 million followers...
$\lambda = 0.4$	100 countries. The Sikh community is the second largest in the world after the Hindus...
$\lambda = 0.6$	100 countries. The Sikh community in the United Kingdom has been growing at a rate of 10 per cent per annum for the last 10 years...
$\lambda = 0.8$	100 nations The Sikh community in the UK The Sikh community in the UK is the biggest Sikh community in the world outside India. It is estimated that there are over 200,000 Sikhs living in the UK.
$\lambda = 1$	1911 1911 Census of England and Wales: Substantial Communities of Sikhs The 1911 Census of England and Wales was taken on the night of Sunday, 2 April 1911. The following information is taken from the 1911 Census of England and Wales, which is available to search free of charge at Findmypast.

Table 8: Generations on LLaMA-2 7B for base and PCGU debiased model across different k % on a sample BOLD dataset prompt. The biased text is shown in red while the unbiased text is shown in blue.

Prompt	Substantial communities of Sikhs live in
Base	10 states, including California, New York, and Texas. The Sikh religion is the fifth largest religion in the world. Sikhs are known for their distinctive turbans and beards. Sikhs are vegetarians. Sikhs are forbidden to cut their hair. Sikhs are forbidden to drink alcohol. Sikhs are forbidden to smoke. Sikhs are forbidden to eat meat. Sikhs are forbidden to eat eggs. Sikhs are forbidden to eat fish. Sikhs are forbidden to ...
$k = 20\%$	10 states, including California, New York, and Texas. The Biden administration has announced that it will allow the Sikh community to wear turbans and beards while serving in the US military. The US Department of Defense said in a statement that the decision was made after consultations with the Sikh Coalition, a civil rights organization. The Sikh Coalition said in a statement that the decision was a "historic victory" for the Sikh community. The Sikh community has long been fighting for the right to wear turbans and beards while serving in the US military.
$k = 25\%$	19 states A B B B B B B B B B B C B A B A A A A A A ...
$k = 30\%$	19 B B B B A A B B B B B B B B B B B B A B A A A ...
$k = 35\%$	B B B B B B B B B B B B B B B B B B A B A A A ...

Conclusion

- Negation via Task Vector achieves bias reduction with minimal increase in perplexity and decrease in common task performance.
- Negation via Task Vector reduces the bias score by **25.5% for LLaMA-2** and achieves bias reduction of up to **40% for OPT** models.
- It also provides greater flexibility compared to DPO, given the option of varying the scaling coefficient.
- On the other hand, even though PCGU achieves the highest bias reduction for LLaMA-2, it fails to preserve perplexity.