FairCon LLM: Improving fairness by leveraging Contrastive learning in LLMs

Sachit Gaudi
Department of Computer Science
Michigan State University
gaudisac@msu.edu

Abstract

Large Language Models (LLMs) have experienced a surge in popularity in recent times, owing to their remarkable ability to follow instructions and demonstrate success across a wide range of Natural Language Processing (NLP) tasks. However, LLMs suffer from a wide range of issues such as harmful generation, fairness, privacy, and robustness. Addressing these issues provides immense value to society and also ensures responsible use of technology.

In this work, we emphasize the existence of fairness-related concerns in large language models (LLMs). Given the significant compute requirements and the discrete nature of LLMs, we are the first to propose a stable adversarial procedure in the context of LLMs. These procedures can be extended to pre-processing techniques, which operate under the assumption of black-box models. We demonstrate that bias in generation is influenced by bias in prompts, providing a basis for the hypothesis that prompt tuning can steer outputs in a fair direction. To achieve this, we introduce a contrastive learning objective and train the network adversarially using Gumbel softmax. We ensure the stability of this training process by implementing Stochastic Weight Averaging (SWA) and address the compute requirements using LoRA adapter. Our findings suggest that the contrastive learning method notably enhances fairness.

1 Introduction

LLMs are deployed in wide array of use cases such as hiring Gan et al. (2024) and healthcare Li et al. (2024), which impact the lives of humans, it puts responsibility on the companies leveraging the technology to comply with regulations. This includes the evaluation and mitigation of bias in hiring decisions based on race and gender. Companies also have an additional responsibility to disseminate harmless, unbiased, and truthful information. Vesnic-Alujevic et al. (2020) calls for AI policy to make companies accountable for privacy, hate speech, and bias.

Work by Wang et al. (2023), show that LLMs suffer wide range of problems such as harmful generation, fairness, privacy, and robustness. the methods mitigating these issues are typically formulated as adversarial tasks, where the goal of the adversary to trigger wrong behaviour and LLMs should be robust to such attacks. In the previous survey paper we Gaudi (2024) have outlined various challenges of adversarial training LLMs, such as discreet text domain and huge compute resources, while also summarizing contributions from researchers aimed at addressing these challenges in the context of harmful content generation. However, these techniques, when directly applied to fairness, they often fail to replicate the success demonstrated for mitigating harm. InstructGPT (Ouyang et al. (2022)) have shown improvements in toxicity over GPT-3 but fails to mitigate bias.

In this work we focus on mitigating bias in LLMs. In Section 1.1 we introduce different notions of fairness. In Section 2, Drawing inspiration from Wang et al. (2023), we evaluate the fairness on

Preprint. Under review.

surrogate task, where the goal is to predict the income based on the different parameters, including gender constructed as text. The advantage of this setting is we can model few shots samples as prompt and control the bias and study the effect of bias on generation. We find that prompt is very critical in controlling the bias. Based on this result we tune the LLM to be fair to the most unfair prompt. This is a adversarial training procedures. The goal of the adversary is to generate most unfair prompt and the goal of LLM is to be fair to adversary, which are extensively studied extensively in Section 3.

1.1 Fairness

The survey Cruz and Hardt (2023) outlines various approaches used to enhance fairness. Fairness in machine learning has branched into three main categories: pre-processing, in-processing, and post-processing. In-processing methods assume access to the complete model, while post-processing methods assume only access to the features, which is not feasible in a black-box setting. However, each method can be extended to the others. For example, if the encoder is frozen and only the classifier is trained, then in-processing techniques can be modified to post-processing. Similarly, if gradients are propagated to the input, the method can be translated into pre-processing.

Figure 1: Causal relation between s, X, and Y.
Left figure case 1: s acts as a confounder, affects $\hat{Y}$ through X and directly
Right figure case 2: $X \perp\!\!\!\perp s$, X and s affect $\hat{Y}$ independently.

Current fairness literature offers multiple definitions of fairness. One such definition is Demographic Parity Difference (DPD), defined as:

$$M_{dpd} = \Big|\Pr(\hat{Y} = 1|s = 0) - \Pr(\hat{Y} = 1|s = 1)\Big| \tag{1}$$

DPD measures the change in model behavior by altering the sensitive attribute while keeping everything else constant. However, this definition fails in Case 1 of figure 1, where controlling for s opens a backdoor path to X. To address this shortcoming, Hardt et al. (2016) proposed Difference in Equalized Odds(DEO), which measures the absolute difference in false positive or false negative rates for all groups. In this paper, we calculate the sum of both and refer to it as DEO:

$$M_{deo} = \sum_{y \in 0,1} |Pr(\hat{Y} = 1|s = 0, Y = y) - Pr(\hat{Y} = 1|s = 1, Y = y)| \tag{2}$$

2 Problem Formulation

We investigate fairness in GPT models, we adopt the framework proposed by Wang et al. (2023). Our task involves leveraging generative models for classification on the Adult dataset. We construct natural language queries from the dataset features and utilize next token prediction to classify whether a person will earn more than $50,000.

GPT models struggle with zero-shot learning in generating meaningful next tokens for the task at hand. To address this limitation, we employ few-shot learning by providing the model with curated samples, guiding it to output binary classifications (1 or 0).

To investigate bias in LLMs we conduct experiments focusing on the Adult Dataset, addressing simplifications for clarity. Recognizing an inherent imbalance in the dataset ($\times 5.23$), we first balance the occurrences of y=1 and y=0. Given the use of a few-shot data points for guiding predictions, the bias introduced by these few-shot samples significantly influences the query bias. We measure bias using bias parity, denoted as b_{P_c}, calculated as $P(y = 1|s = 0) - P(y = 1|s = 1)$. Here, s represents the sensitive attribute (gender in our example), and y indicates income status, where 1 denotes income greater than $50K$, and 0 denotes income less than $50K$. Control over b_{P_c} is achieved by sampling 200 data points according to the specified distribution.

A natural bias of $b_{P_c} = 0.1312$ is present in the dataset. When we randomly sample from the dataset, the few-shot samples inherit the same bias. To adjust the sampling, we can independently sample from $P(y = 1|s = 0)$ and $P(y = 1|s = 1)$. Alternatively, we can employ the counterbalance technique to make the prompt fair by creating a sample where gender is switched from male to female while keeping other attributes constant. This modification results in a bias parity of 0 for the prompt.

b_{P_c}	ACC ↑	M_{dpd} ↓	M_{eod} ↓
0.00	75.5	**0.0049**	**0.0083**
0.13	**85.0**	0.0080	0.0180
0.50	70.5	0.0411	0.0429
1.00	68.5	0.0940	0.1019

Table 1: Few shot (16) performance of GPT models under different bias of the prompts

b_{P_c}	ACC ↑	M_{dpd} ↓	M_{eod} ↓	Counterbalance
0.00	81.5	**0.0028**	**0.0082**	✓
0.00	75.5	0.0049	0.0083	×

Table 2: Counterbalance by adding conuterfactuals

From Table 1 and Table 2, it is evident that the selection of few-shot samples significantly influences the generation outcome. By managing the bias in the prompt, we can regulate the fairness of the model. However, we also observe a decrease in accuracy when deviating from the inherent bias present in the dataset. Thus, there exists a trade-off between fairness and accuracy.

Having demonstrated that bias in the prompt influences to the generated output, Instead of manually controlling the prompt we will influence the training with by characterising fairness with prompt.

3 Method

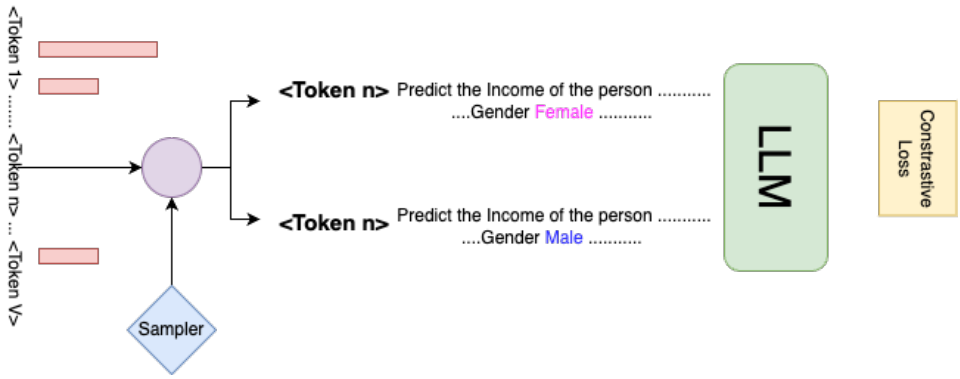


Figure 2: Contrasting training procedure where the LLM and the token probabilities are trained alternately. The objective is to identify biased prompts, with the LLM subsequently fine-tuned to be robust against such prompts.

3.1 Bi-level adversarial optimisation

The above idea can be formulated as a the objective below

$$\begin{aligned} &min_\theta min_\phi \quad \mathcal{L}(f_\theta(g_\phi, x_i)) + \lambda ||f_\theta(g_\phi, x_{mi}) - f_\theta(g_\phi, x_{fi})||_2 \\ &\textbf{subject to } g_\phi = arg\max ||f_\theta(g_\phi, x_{mi}) - f_\theta(g_\phi, x_{fi})||_2 \end{aligned} \tag{3}$$

$\mathcal{L}$ is Task Loss, here predicting if the income is grater than \$50K and g_ϕ is the maximum unfair prompt, which is a adversarial network. and f_θ is the LLM from which we want to remove bias.

The Equation 3 is the formulation of bi-level optimisation problem. where the goal of LLM is to finetune on the task and the goal of the adversary is to find the prompt that makes the output that prompts the LLM to leverage gender information.

We can solve this optimisation by ADMM Boyd et al. (2011), alternating between optimisation of θ and then ϕ till the convergence.

3.2 Gumbel Soft-max Reparametrisation

The gumbel softmax trick is given by sampling in the forward pass $f(v) = arg\max_{\mathcal{V}}[v_1 \quad v_2 \quad v_3]$ and the backward pass propagates gradients as if function, f is replaced with simple softmax function $\frac{\partial f(v)}{\partial v} = \sigma(1-\sigma), where\ \sigma = Softmax([v_1 \quad v_2 \quad v_3])$. It can be derived as a reparametrisation trick Jang et al. (2016). This trick enables gradients to propagate back to inputs, allowing adversarial techniques designed for continuous domains to be adapted for text.

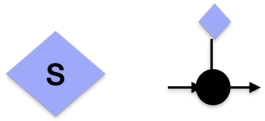


Figure 3: Re-parametrisation trick

3.3 Extension to prompt tuning

$$min_{\phi}||f_{\theta}(g_{\phi}, x_{mi}) - f_{\theta}(g_{\phi}, x_{fi})||_2 \tag{4}$$

If we have no access to the model weights. We can only tune the prompt according to Equation 4, which is equivalent to keeping the LLM weights frozen and solving one part of the bi-level optimisation in Equation 3.

4 Experimental Setup

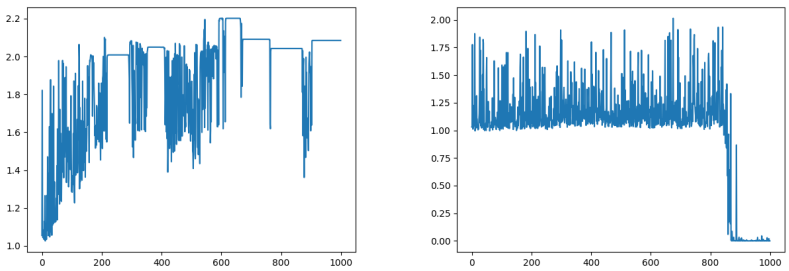
We will restrict ourselves to the simplest setting of improving the fairness of LLM, here we consider 1.3B GPT-neo and also the restrict adversary to the categorical distribution. This simplest setting is to prove the gradient propagation. We can however replace the categorical distribution to another LLM. In this work, we perform experiments to show the improvements in fairness on the toy setup mentioned in the Section 2. The code is available at [1]

4.1 LoRA Adapter

Computing gradients for 1.3B parameters, and having the network computational graph on memory is memory intensive. However, we leverage LoRA adapter Hu et al. (2021) to train only the query and value projection matrices of the transformer block. thereby reducing the total trainable parameters to 0.5% .

4.2 Stochastic Weight Averaging

Gumbel softmax is a stochastic process, therefore training is highly unstable. Some form of weight averaging Karras et al. (2023) is suggested to improve the stability of training. Here we use SWA Izmailov et al. (2018) to average the model weights, which will improve generalisation and also we see the stability at the end of the training and also reduces influence on learning rate.



(a) Without SWA (b) SWA

Figure 4: Training a known distribution with gumbel softmax

[1] https://github.com/sachit3022/FairCon

5 Results

Training	ACC ↑	M_{dpd} ↓	M_{eod} ↓
Fine tuned LLM	63.0	0.183	0.142
Fine tuned LLM with fairness constraints	51.0	0.136	0.081

Table 3: Training LLM with an implicit bias of b_{P_c} = 0.5 with and without fairness constraints.

In Table 3, we show that training with fairness constraints will results in improvement in fairness. We examine the sample generated by the adversarial network as a prompt. Here is an example: "aerial 134 Att claims Generic execute Whatever rink reservoirs Dragon." Since we have not imposed any constraints on the prompt, it lacks meaningful language structure. However, by LLM, instead of categorical distribution, we can observe prompts that resemble human language through sampling.

6 Conclusion

There are no papers for training LLMs with adversarial objectives. This paper by Ganguli et al. (2022) predicts that end-to-end adversarial min-max training with RL will lead to superior results, although the stability of RL remains a concern and may result in model collapse but there is currently a lack of empirical evidence demonstrating its success. We are the first one to study adversarial optimisation in the context of LLM.

We introduce the stability to the training process by introducing SWA.

We reinforce that bias in the prompt can be translated to the bias in generation, we therefore make the LLM fair to the unfair prompt, thereby making the LLM robus to prompt based attacks.

7 Related Works

In recent times, there has been significant interest in the field of fairness, as outlined in the survey by Caton and Haas (2020), which discusses various risks associated with unfair models and highlights the direction of fairness research. Additionally, the work of Dehdashtian et al. (2024) has extended fairness techniques to a multi-modal setting. However, fairness remains a relatively understudied area in generative models. Current techniques to address fairness in large language models (LLMs) primarily rely on prompt tuning and pre-processing methods. Some approaches involve manually crafting prompts, as mentioned in Si et al. (2023), while others leverage training an LLM to automatically adjust prompts using techniques such as Gumbel softmax Xu et al. (2023) or methods proposed by Wu et al. (2024), which build upon the success of instruction tuning Ouyang et al. (2022) to find prompts that yield fair outputs for a center class of instructions. However, due to the unstable optimization procedures of techniques like Gumbel softmax or Proximal Policy Optimization (PPO), they are not easily extended to adversarial training.

8 Future works

In this we show that gradient based techniques are still applicable and the techniques that are designed for the Continuous domain can be adapted to the LLMs. However, we need to investigate the stability of the training procedure and demonstrate the success on a large dataset.

As the current direction of LLM is black-box accuess, We need to investigate the transfer learning capabilities of such models when applied to commercial LLMs like ChatGPT.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *ACM Computing Surveys*.

Cruz, A. F. and Hardt, M. (2023). Unprocessing seven years of algorithmic fairness.

Dehdashtian, S., Wang, L., and Boddeti, V. N. (2024). Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. *arXiv preprint arXiv:2403.15593*.

Gan, C., Zhang, Q., and Mori, T. (2024). Application of llm agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315*.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gaudi, S. (2024). All things adversarial in llms: A survey.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. (2023). Analyzing and improving the training dynamics of diffusion models. *arXiv preprint arXiv:2312.02696*.

Li, J., Dada, A., Puladi, B., Kleesiek, J., and Egger, J. (2024). Chatgpt in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, page 108013.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., and Wang, L. (2023). Prompting gpt-3 to be reliable.

Vesnic-Alujevic, L., Nascimento, S., and Pólvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on european policy frameworks. *Telecommunications Policy*, 44(6):101961. Artificial intelligence, economy and society.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., and Li, B. (2023). Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wu, Z., Gao, H., Wang, Y., Zhang, X., and Wang, S. (2024). Universal prompt optimizer for safe text-to-image generation. *arXiv preprint arXiv:2402.10882*.

Xu, H., He, P., Ren, J., Wan, Y., Liu, Z., Liu, H., and Tang, J. (2023). Probabilistic categorical adversarial attack and adversarial training. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38428–38442. PMLR.