INDUCING GROUP FAIRNESS IN PROMPT-BASED LANGUAGE MODEL DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifiers are used throughout industry to enforce policies, ranging from the detection of toxic content to age-appropriate content filtering. While these classifiers serve important functions, it is also essential that they are built in ways that minimize unfair biases for users. One such fairness consideration is called *group fairness*, which desires that different sub-population of users receive equal treatment. This is a well-studied problem in the context of 'classical' classifiers. However, the emergence of prompt-based language model (LM) decision making has created new opportunities to solve text-based classification tasks, and the fairness properties of these new classifiers are not yet well understood. Further, the 'remediation toolkit' is incomplete for LM-based decision makers and little is understood about how to improve decision maker group fairness while maintaining classifier performance. This work sets out to add more tools to that toolbox. We introduce adaptations of existing effective approaches from the classical classifier fairness to the prompt-based classifier space. We also devise simple methods that take advantage of the new structure of prompt-based decision makers and operate at the prompt level. We compare these approaches empirically on real data. Our results suggest that adaptations of approaches that are effective for classical classifiers remain effective in the LM-based classifier environment. However, there is room for further exploration of prompt-based remediation methods (and other remediation methods that take advantage of LM structure).

1 INTRODUCTION

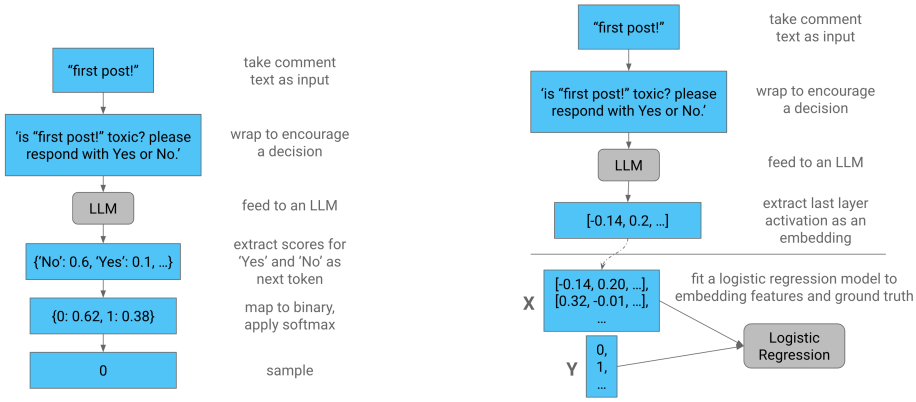
Language models (LMs) have shown impressive performance across many tasks and are now being deployed across high-stakes applications such as financial Wu et al. (2023) or medical Singhal et al. (2023) domains. In particular, zero-shot LM-based classifiers Wei et al. (2022a); Anil et al. (2023) have achieved state-of-the-art performance on several natural language classification benchmarks and are being widely adopted for decision making. More recently, such classifiers are leveraged as a reward signal to align models with AI feedback Bai et al. (2022). Hence, it is important to ask: *How fair are the classification decisions made by LMs?*

In this paper, we consider two classes of LM-based classifiers: (i) prompted ("out-of-the-box") LM classifiers and (ii) trained classifiers on top of last-layer embeddings extracted from an LM. We first assess whether these LM-based classifiers satisfy a widely adopted classifier group fairness notion called equal opportunity (EO) Hardt et al. (2016); Prost & Beutel (2020). EO is measured as the difference between the false positive rates (FPR) of different demographic groups, where negative outcome is considered an advantaged class. For example, consider a toxicity detection classifier where being labeled as toxic leads to some content moderation policy. It is therefore desirable for content from all demographics to be falsely marked as toxic with an equal rate. We find that prompted LM classifiers demonstrate a significant gap in FPR across multiple demographic groups in the Civil Comments toxicity detection benchmark Borkan et al. (2019), with Muslim and Jewish groups having 89% and 48% higher FPR as compared to the Christian group. The gap is further increased when we compare embedding-based classifiers with Muslim and Jewish groups having 124% and 71% more FPR compared to the majority group.

We then benchmark the effectiveness of two types of group fairness remediation techniques: (i) prompting-based and (ii) regularization-based remediation methods. For prompting-based methods,

we study the effectiveness of different group-agnostic and group-aware fairness encouraging natural language prompts. In the context of regularization-based methods, we study an post-processing remediation method (Ţifrea et al., 2024) and in-processing method (Prost et al., 2019; Beutel et al., 2019). We find that prompt-based remediation methods are unable to decrease the FPR gap in our experiments - with Muslim and Jewish groups still having FPR about 40% higher than the Christian group.

Contributions. (1) We assess the group fairness of two classes of LM-based classifiers (i.e., prompt-based and embedding-based) and show that they do not satisfy equal opportunity (EO) along identity aspects such as religion, race, ethnicity, sex. (2) We evaluate three different remediation techniques (i.e., prompting, in-processing, and post-processing) within the two studied classes of LM-based classifiers. We find that prompting-based remediations fail to achieve lower false positive rates, and that regularization-based approaches achieve better fairness-performance tradeoffs across both classifier classes. (3) We find that in-processing remediation achieves better fairness-performance trade-offs than post-processing methods, but may not be always a feasible option due to limited access to the internals of the model.



(a) **Prompt-based Classifier** (b) **Embedding-based Classifier**

Figure 1: Classification flow diagrams for prompt-based and embedding-based classifiers. Decisions are encouraged via 'text wrappers' that nudge the LM to make a classification decision. (1a) For prompt-based classifiers, we treat the wrapped text as a prefix and query the LM for two postfix tokens (such as 'Yes' or 'No') that represent positive and negative decisions. We apply a softmax to these scores to obtain a probability distribution over the classification result and use this for decision making. (1b) For embedding-based classifiers, we assume that the LM is 'introspective' and can supply its activations. We instead query the LM for the activations of its last layer to serve as an embedding. We collect those embeddings into a design matrix then fit a logistic regression model on that matrix and corresponding labels. The logistic regression model is then used for downstream decision making.

2 Problem Setup

Equal opportunity notion of group fairness. Consider a binary classifier where the positive outcome is considered to be a disadvantaged class, e.g., being detected as toxic. The widely adopted equal opportunity (EO) notion of fairness requires that different demographics are disadvantaged with the same rate, i.e., the false positive rate (FPR) of the classifier is equal for all demographics (Hardt et al., 2016; Prost & Beutel, 2020; Lowy et al., 2022; Hort et al., 2022). For example, this means that the content from different demographics is desired to be falsely marked as toxic at an equal rate.

LM-based classifiers. We explore two classes of classifiers: (i) prompt-based LM classifiers, and (ii) trained classifier on top of extracted last-layer embeddings of LMs. The former does not involve any training and leverages the instruction following ability of LMs. In contrast, the latter requires supervised training of a classifier, and is also commonly referred to as a trained linear head on top of the last-layer representation. In both cases, we assume *grey-box* access to the LM, i.e., access to the logits (for prompt-based classifiers) or last layer activations (for embedding-based classifiers) produced by the language model prior to sampling.

- **Prompt-based LM classifier.** Fig. 1 gives a schematic diagram of the classification methodology described in Sec. 2. Note that we use the following wrapper text for our toxicity classifier: We prompt an LLM with the toxicity classification task formulated as yes or no question. By assessing the log-probabilities of the "Yes" and "no" tokens and applying a softmax function, we are able to extract a distribution over negative and positive classes. A schematic diagram of the process is given in Fig. 1a and our prompt is given below.

  > {TEXT_TO_CLASSIFY} Categorize the comment on the basis of toxicity. If the comment is toxic (harmful, malicious, derogatory, threat, insult, identity attack, etc.) output Yes, otherwise output No.

- **Embedding-based classifiers.** We prompt an LLM in the exact same way as above, but extract the final-layer activations of an LM for a training set. Using the collected pairs of (activation, label), we then train a logistic regression model to predict the ground truth labels. During inference time, we obtain the last layer embedding of the wrapped classification input (see Figure 1b) and feed it into the trained classifier. This approach can be thought of as a form of parameter efficient supervised fine-tuning. A schematic diagram is given in Fig. 1b in the appendix.

3 Experiment Setup

Models. We use PaLM2 S Anil et al. (2023) as the base model, and PaLM 2L in a model transfer experiment.

Dataset & Tasks. We use the Civil Comments Identity Borkan et al. (2019) dataset as it provides the information necessary to evaluate and remediate both group fairness and performance. For processing details, please see Appendix D.

Evaluation metrics. We quantify the equality of opportunity *fairness* for each group as compared to the majority group. One metric for achieving this is the False Positive Rate ratio (FPR ratio) between the group and the corresponding majority:

$$FPR\ Ratio = \frac{FPR_{group}}{FPR_{majority}}. \tag{1}$$

Ideally, we want the FPR ratio to be close to one. To measure *performance* of the classifiers, we report Area Under the Receiver Operating Characteristic Curve (ROC AUC) over the test split which includes all demographic groups.

4 Remediation Methodology

In this section we describe our prompt-based remediation methodology and our adaptation of two 'classical' remediation methods to the LLM environment. This is followed by an empirical comparison of the methods.

4.1 Prompting

We explore the performance and fairness of three prompt-based methods of increasing specificity. Using a running example of remediating with respect to the Jewish group, we have:

- **Please Be Fair (`PBF`)**: 'Please be as fair as possible when making a decision' is appended to the prompt.
- **Please Be Fair to Super Group (`PBF2SG`)** 'Please be as fair as possible when making a decision about comments about religious groups or that mention religion' is appended to the prompt.
- **Please Be Fair to This Group (`PBF2TG`)** 'Please be as fair as possible when making a decision about comments that mention Judaism or Jewish people' is appended to the prompt.

This is a particularly challenging environment for prompt-based methods because we are interested in inducing group fairness, a subpopulation-level behavior, but apply the same prompt to each instance.

This is also difficult to define an in-context method because it does not make sense to present an instance as being group-fair or not.

Note also that, while the methods described in the next sections provide a hyperparameter that can tune the fairness versus performance trade-off, prompt-based methods have no such capability.

4.2 IN-PROCESSING

Let (x, y) represent an input prompt and a label, where $x \in \mathcal{X}$, and $y \in \{0, 1\}$. We use the loss:

$$L_{IP} = L_{CE}(\hat{Y}, Y) + \lambda D\left(\hat{Y}; G|Y=0\right), \tag{2}$$

Where L_{IP} is the in-processing loss, L_{CE} is the usual cross entropy loss for learning a classification task, and D is a statistical divergence promoting the decision $\widehat{Y}$ to be independent of the sensitive attribute G.

Among many choices for the statistical divergence, we focus on Min Diff Prost et al. (2019), which is an approach that has been successful in remediating to achieve equality of opportunity in the 'classical' setting in industrial settings Prost & Beutel (2020). The central insight behind the Min Diff approach is that the distance between the probability of a false positive for instances from groups and majority can be included as the loss. This encourages those distributions to be closer together, which in turns pushes the false positive rates for group and majority towards each other. In particular, Min Diff uses *MMD* as a Maximum Mean Discrepancy kernel that gives the distance between the distributions of probability of a false positive for group and associated majority, and λ is a parameter that trades off between the two loss terms (and thus between performance and fairness).

To adapt Min Diff to the LM decisions, we use the Min Diff loss during fine tuning. This approach cannot work in the zero-shot case where no fine tuning is performed.

4.3 POST-PROCESSING

Recent work has demonstrated how in-processing techniques can be adapted to post-processing scenarios Ţifrea et al. (2024). We leverage this approach to fit a post-processed 'emfairening' model. The emfairening model's predictions are added to the unremediated models predictions in logit space, i.e., for all prompts x and label y:

$$\pi_{pp}(y|x) \propto \pi_{\text{ref}}(y|x)\pi_{\text{emf}}(y|x), \tag{3}$$

where $\pi_{\text{ref}}(y|x)$ is the baseline prediction distribution, $\pi_{\text{emf}}(y|x)$ is the emfairening model's distribution, and $\pi_{pp}(y|x)$ gives the combined post-processed distribution. The emfairening model can be trained with the following loss:

$$L_{PP} = KL\left(P_{pp} \| P_{ref}(y|x)\right) + \lambda D\left(\hat{Y}; G|Y=0\right) \tag{4}$$

where *KL* is the Kullback-Leibler divergence and D is a fairness promoting divergence. In this paper, We use the industry trusted *MMD* kernel Prost & Beutel (2020). The KL divergence term prevents 'catastrophic forgetting' in the emfairening model; that is, we encourage the emfairening model to not stray too far from the performant baseline model. This ensures that we maintain acceptable classifier performance when making emfairened predictions. As with the in-processing method, λ trades off between the two loss terms and thus dials between fairness and performance.

We note that the post-processing formulation here resembles controlled decoding methods used to steer the generation of language models towards high reward outcomes Mudgal et al. (2024) with a reward that captures the fairness of the outcome.

We also note that we are free to choose the baseline model $\pi_{\text{ref}}(y|x)$. As such, we can apply this approach directly to any LM in a zero-shot manner or apply it to a fine-tuned model.

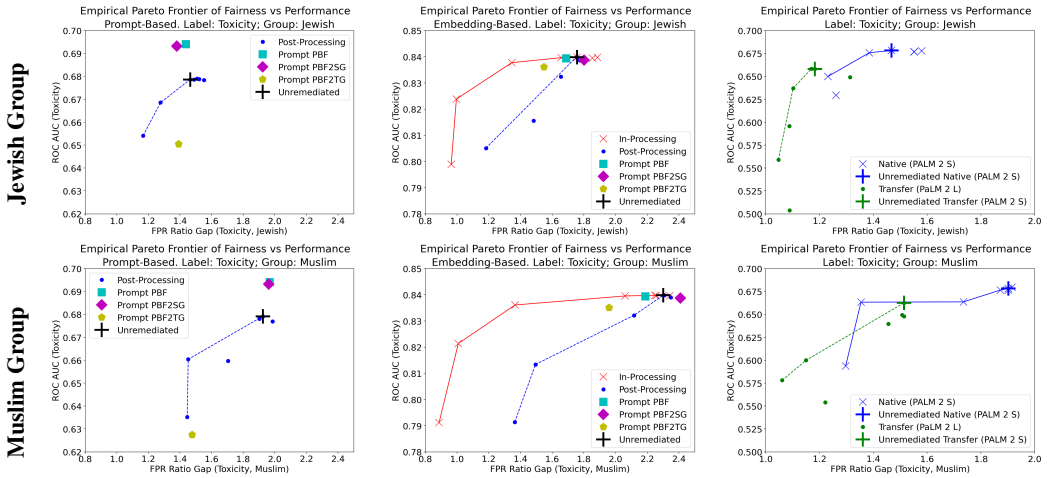
5 EXPERIMENT RESULTS

Group fairness without remediation. First, we evaluate the two classifier approaches on PaLM 2 S Anil et al. (2023) with respect to equality of opportunity. These experiments follow the methodology described in Sec. 2 and use the Civil Comments Identity Borkan et al. (2019) dataset.

Group	Prompt-based Classifier	Embedding-based Classifier
Muslim	1.89	2.24
Jewish	1.48	1.71

Table 1: False positive rate ratios that quantify the magnitude of EO violation for a classifier built on PaLM 2 S. We only include the two groups with the highest gaps. Here we used 'Christian' as the majority group.

The FPR ratios for the two groups with the highest ratio gaps are given in Tab. 1 and the results for all groups are given in Tab. 2 (Appendix). These elevated FPR ratios imply a need for group fairness remediation. In the remainder of this section, we will focus on remediation approaches and analyze their empirical performance vs fairness trade-offs.



(a) **Prompt-based Classifiers** (b) **Embedding-based Classifiers** (c) **Model Transfer**

Figure 2: Pareto frontiers of different remediation techniques. The left plot shows the performance and fairness of prompt-based classifiers, and the middle plot of embedding-based classifiers. The unremediated classifier setting is denoted by a '+' and prompting-based remediation methods are denoted by single symbols. Note that the in-processing baseline is inapplicable to prompt-based classifiers. Each point for in-processing and post-processing is generated by setting different values for λ in Equations (2) and (4) in the appendix. The dashed and solid lines give the Pareto frontier where performance can only be gained by sacrificing fairness, for post-processing and in-processing, respectively. The right plot gives the effect of model transfer. We fit a post-processing remediation model to the PaLM 2 S model then compare the effects of applying it to PaLM 2 S (native) versus the larger PaLM 2 L model (transfer). The lines give the Pareto frontier (solid for native and dashed for transfer).

Remediation of prompt-based classifiers We start with analyzing remediation techniques in prompt-based classifiers. Results are shown in Fig. 2a. For the post-processing method, each point in the plot is generated by varying the regularizer strength (see Appendix 4.3 for more details).

We observe that the post-processing method improve fairness without severely degrading the performance of the classifier. In contrast, prompt-based remediations either increase performance while keeping the high FPR ratio, or reduce the FPR ratio at the cost of performance.

Remediation of embedding-based classifiers. We compare the Pareto frontiers of fairness and performance for each remediation technique in Fig. 2b. As before, each point of regularization-based techniques in the plot is generated by varying the regularizer strength that trades off between performance and fairness terms in the objective function.

There are a few takeaways from these experiments. First, embedding-based classifiers show superior performance compared to prompt-based classifiers (see Fig. 2a). Second, as before, we observe improved group fairness for regularization-based methods without significantly degrading the per-

formance of the classifier. However, the in-processing technique generally performs better than the post-processing technique in this setting. Finally, prompt-based remediation methods show some fairness and performance benefit but are generally less controllable and less effective than in-processing and post-processing methods.

Transfer of remediation to an unseen model. In this experiment, our goal is to apply the remediation to a grey-box model which only provides access to logits (and not the last layer activations), i.e., a prompted LM. As such, we use the Google News 128-dimensional embedding model Bengio et al. (2000)[1] to embed the query and use these embeddings for any subsequent remediation. Our setup enables us to operate in environments where drawing embeddings from the LM is not feasible. One interesting case is where we train a post-processing remediation model on one LM, and then apply it to remediate another LM (with grey-box access); can we reuse the existing fairness model to improve fairness with the new model?

Fig. 2c gives the results this model transfer learning scenario. Note also that the Pareto frontier achieved by the smaller PaLM 2 S model in Fig. 2c, where Google News embeddings are used, is just slightly degraded from the Pareto frontier given in the upper plot of Fig. 2a where model activiations are used as embeddings, making this a promising approach for fairness remediation. Importantly, we find that post-processing model is still able to improve fairness when transferred (although this comes at the cost of higher performance degradation than when applied to the model that it was trained on). This suggests that we may be able to train universal fairness mitigation heads that could be applied to any LM with grey-box access and provide fairness benefits for classification tasks.

6 Concluding Remarks

Fairness is an important consideration for classifiers in industry. Even though the research community has made significant progress on training fair classifiers, the recent shift towards prompt-based classifiers requires exploring new fairness solutions for prompt-based decision making.

We study the group fairness of two classes of LM-based classifiers. We identify that LM-based classifiers may exhibit group unfairness. We introduce and evaluated three remediation techniques to improve fairness while maintaining acceptable performance for LM-based classifiers. We find that prompt-based techniques offer limited benefit and are in general outperformed by in-processing and post-processing techniques.

Within the scope of the evaluated prompts, we find that embedding-based classifiers exhibit superior performance compared to "out-of-the-box" prompt-based classifiers in our setup. In addition, we find that the in-processing method consistently provides favorable performance vs fairness trade-off on embedding-based classifiers. We conclude that for remediating an embedding-based classifier, in-processing is a more robust approach. In other LM-based classification settings where in-processing cannot be applied (prompt-based classifiers and transfer tasks) the post-processing technique provides promising results.

We find that the prompt-based remediation methods have little to no impact of prompts on fairness, while counter-intuitively, we observe that fairness-oriented prompts may slightly improve performance in some cases for the less specific 'Please be Fair' (`PBF`) and 'Please be Fair to Super Group' (`PBF2SG`) methods. This is not surprising given that fairness is a distributional issue, and hence prompting may not necessarily provide the distribution matching effects that we expect from remediation.

References

Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1418–1426, 2019.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

[1] `https://www.kaggle.com/models/google/nnlm`.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. Your fairness may vary: Pretrained language model fairness in toxic text classification. *arXiv preprint arXiv:2108.01250*, 2021.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 453–459, 2019.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Evgenii Chzhen and Nicolas Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *arXiv preprint arXiv:2007.14265*, 2020.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1236–1270, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.88. URL https://aclanthology.org/2023.findings-emnlp.88.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295.

Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural Réyni minimization for continuous features. *arXiv preprint arXiv:1911.04929*, 2019.

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via Rényi minimization. *arXiv preprint arXiv:2009.03183*, 2020.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pp. 869–874. IEEE, 2010.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10383–10405, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.643. URL https://aclanthology.org/2023.emnlp-main.643.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pp. 6565–6576. PMLR, 2021.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization. *Transactions of Machine Learning Research*, 2022.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. Controlled decoding from language models. *International Conference on Machine Learning*, 2024.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Flavien Prost and Alex Beutel. Mitigating unfair bias in ML models with the MinDiff framework. Google Research Blog, 2020. URL https://research.google/blog/mitigating-unfair-bias-in-ml-models-with-the-mindiff-framework/.

Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

Edward Raff, Jared Sylvester, and Steven Mills. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 243–250, 2018.

Goce Ristanoski, Wei Liu, and James Bailey. Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 1529–1532, 2013.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022a.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022c.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

Ruicheng Xian and Han Zhao. Optimal group fair classifiers from linear post-processing. *arXiv preprint arXiv:2405.04025*, 2024.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *International Conference on Machine Learning*, pp. 37977–38012. PMLR, 2023.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Alexandru Ţifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. FRAPPÉ: A group fairness framework for post-processing everything. *International Conference on Machine Learning*, 2024.

A Related Work

While LMs have been broadly studied in the generative case for trustworthiness, e.g., diversity, stereotypes, gender bias, and toxicity Nadeem et al. (2021); Liang et al. (2021); Deshpande et al. (2023); Lahoti et al. (2023), their fairness in classification problems remains under-explored. Tamkin et al. (2023) provided a method for evaluating how biased a language model may be by generating hypothetical prompts with group information and making decisions by fitting a mixed effects model. The authors of the Flan-T5 model Chung et al. (2022) published group-level performance of a toxicity classifier fit to the Civil Comments Identity Borkan et al. (2019) dataset. Baldini et al. (2021) explored remediation methods for achieving equalized odds for different embedding-based classifier models. To our knowledge, this is the first paper that proposes and empirically evaluates methods for remediating LM-based classifiers drawn from LLMs with respect to equal opportunity fairness.

There is also rich literature on fairness in classical (i.e. non-LM-based) classification. In this paper we focus on the equal opportunity (EO) notion of group fairness Hardt et al. (2016); Prost & Beutel (2020) which is achieved for a group and classifier when the false positive rate (or false negative rate) of the classifier is the same for instances drawn from that group when compared with instances drawn from the majority group.

Methods for improving group fairness can generally be categorized in three main classes: *pre-processing*, *in-processing*, and *post-processing* methods. Pre-processing algorithms (Feldman et al., 2015; Zemel et al., 2013; Calmon et al., 2017) transform the biased data features to a new space in which the labels and sensitive attributes are statistically independent. In-processing methods (Kamiran et al., 2010; Ristanoski et al., 2013; Quadrianto & Sharmanska, 2017; Zafar et al., 2017; Berk et al., 2017; Donini et al., 2018; Raff et al., 2018; Aghaei et al., 2019; Prost et al., 2019; Beutel et al., 2019; Grari et al., 2019; Taskesen et al., 2020; Grari et al., 2020; Cho et al., 2020; Chzhen & Schreuder, 2020; Jiang et al., 2020; Lowy et al., 2022) add a regularizer or constraint to the learning objective. Post-processing approaches (Hardt et al., 2016; Pleiss et al., 2017; Alghamdi et al., 2022; Xian et al., 2023; Xian & Zhao, 2024; Ţifrea et al., 2024) improve group fairness properties by altering the final decision of the classifier. See the survey paper by Hort et al. (2022) for a more comprehensive literature survey.

Among all these classical approaches, we believe post-processing approaches are the most compatible for the LM-based decision making. Having said that, most post-processing approaches require access to demographic labels at test time, which is infeasible, especially for LM-based classifiers. That is why we only focused on FRAPPÉ (Ţifrea et al., 2024) which works without access to demographic labels.

B Limitations

We would like to mention a few limitations to our work that could also be seen as opportunities for future work:

- We find that prompting-based remediation methods are less flexible and effective than in-processing and post-processing methods. However, we do not make an exhaustive search of possible prompts and other researchers may find prompting-based remediation methods that work. Furthermore, it has been observed that the capabilities of language models improve with the model size Wei et al. (2022b), and this could have a beneficial effect on prompt-based method effectiveness as LMs become larger and more capable.
- Apart from prompting-based remediation methods, the proposed remediation techniques require grey-box access to the logits of the model prior to sampling, and may not be applicable if the model only provides black-box access.
- Our experiments are focused on equal opportunity (EO) notion of group fairness. There is no guarantee that they will generalize to other notions of fairness, and, importantly, their application does not imply that a classifier is abstractly *fair*, as all different notions of group fairness have their own limitations and might even be at odds with each other.
- LM-based classifier inference is very expensive when compared to simpler models, and the performance of LM-based classifiers does not yet justify that cost (for example, our LM-based classifiers are less effective than baseline methods given by the authors of the Civil Comments dataset paper Borkan et al. (2019)). An implicit assumption of this work is that the performance

of LM-based classifiers will improve enough over time to justify their high inference costs and become deployed systems where fairness considerations are in play.
- We considered only one language model (PaLM 2) and one dataset (Civil Comments Identity) in English. So it remains to be seen how much our findings generalize. Having said that, given that we already find performance disparities across subgroups in this limited case, the need for developing fairness remediation techniques for LM-based decision making systems is justifiably real.
- We only experiment with a few-handcrafted prompts for classification, and did not compare against chain-of-thought Wei et al. (2022c), self-consistency Wang et al. (2022), and automated prompt generation Gao et al. (2021) techniques as adapting them to induce group fairness was not trivial and is left for future work.
- We do not benchmark other popular techniques, such as low-rank adaptation Hu et al. (2021), prompt-tuning Lester et al. (2021), and other parameter-efficient fine-tuning techniques Liu et al. (2022) for the in-processing method.

C Risks

There are three risks we would like to call out:

- Group fairness remediation improves group fairness on a training set. This may fail to generalize to a held-out set under some circumstances (for instance, if there is distributional shift).
- Group fairness remediation improves only group fairness. Importantly, it does not guarantee improvement in other notions of fairness or make a classifier abstractly 'fair.'
- Group fairness remediation methods could be reversed by a malicious actor to worsen the group fairness of a classifier.

D Data and Data Processing

Experiments in this paper are based on the Civil Comments Identity Borkan et al. (2019) dataset. This dataset was selected because it provides the information necessary to evaluate and remediate both group fairness and performance; that is, textual data and several moderation-based labels that classifiers can be trained on and group data that can be used for evaluation and remediation with respect to group fairness.

The dataset contains 405,130 training instances, 21,293 validation instances, and 21,577 test instances. We make use of all three splits in our work. The training set was used for training, validation set for classifier threshold selection, and the test set for reported results.

The Civil Comments identity label and group data are represented as the proportion of raters who believe that a given text instance is an example of various moderation labels as well as various group labels. Note that the group labels correspond to the whether the content of the text is relevant to that group. Because we require binary label and group data for both remediation and evaluation, we treat any non-zero proportion of raters as a positive instance and zero values as negative instances.

E Full Benchmark Ratio Gap Results

We only report the ratio gaps for the two groups with the highest gaps (Jewish and Muslim) in Tab. 1 in Sec. 5 for the unremediated case. The full table is given in Table 2

F Use of Scientific Artifacts

We make use of two scientific artifacts: PaLM 2 Chung et al. (2022) and the Civil Comments Identity Borkan et al. (2019) dataset. Our use of PaLM 2 is consistent with the publication guidelines of the model creators. Our use of Civil Comments is consistent with the 'Public Domain (CC0)' license under which it is released.

Group	Prompt-based Classifier (FPR Ratio)	Embedding-based Classifier (FPR Ratio)
muslim	1.89	2.24
jewish	1.48	1.71
other religion	1.40	1.32
hindu	1.39	1.46
transgender	1.24	1.63
female	1.11	1.05
black	1.06	0.90
asian	0.95	0.36
latino	0.92	0.50
other race or ethnicity	0.91	0.44
homosexual gay or lesbian	0.90	1.13
other sexual orientation	0.86	0.64
buddhist	0.75	1.08
bisexual	0.72	0.77
other gender	0.57	0.85

Table 2: False positive rate ratios that quantify the magnitude of violation of equality of opportunity for an unremediated classifier built on PaLM 2 (Anil et al., 2023)