

# Inducing Group Fairness in LLM-Based Decisions

James Atwood   Preethi Lahoti   Ananth Balashankar   Flavien Prost   Ahmad Beirami  
Google DeepMind

## Abstract

Prompting Large Language Models (LLMs) has created new and interesting means for classifying textual data. While evaluating and remediating group fairness is a well-studied problem in classifier fairness literature, some *classical* approaches (e.g., regularization) do not carry over, and some new opportunities arise (e.g., prompt-based remediation). We measure fairness of LLM-based classifiers on a toxicity classification task, and empirically show that prompt-based classifiers may lead to unfair decisions. We introduce several remediation techniques and benchmark their fairness and performance trade-offs. We hope our work encourages more research on group fairness in LLM-based classifiers.

## 1 Introduction

Large Language Models (LLMs) have shown impressive performance across tasks and are now being deployed across many high-stakes applications such as financial (Wu et al., 2023) or medical (Singhal et al., 2023) domains. In particular, zero-shot LLM-based classifiers (Wei et al., 2022a; Anil et al., 2023) have been shown to achieve state-of-the-art performance on several natural language classification benchmarks and are being widely adopted for decision making. More recently, such classifiers are even used as a reward signal to align the model to AI Feedback (Bai et al., 2022). Hence, it is important to investigate the fairness of such classifiers to different demographics.

While LLMs have been broadly studied in the generative case for 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. In this paper, we investigate specifically if zero-shot and few-shot classifiers satisfy group fairness notions such as Equality of Opportunity (EO) (Hardt et al., 2016), measured as the

difference of false positive rates (FPR) between demographic groups.

Not surprisingly, we find that the prompted zero-shot LLM-based classifiers demonstrate a significant gap in FPR across multiple demographic groups, with Muslim and Jewish groups having 89% and 48% higher FPR as compared to the Christian group in the Civil Comments toxicity detection benchmark (Borkan et al., 2019). The gap is further increased when we compare few-shot prompting based LLM classifiers with Muslim and Jewish groups having 124% and 71% more FPR compared to the majority group.

Fairness mitigation for LLM-based classifiers is challenging due to lack of access to pretraining data and procedures. In addition, little to no fine tuning is usually done in zero-shot and few-shot classification tasks, which makes it hard to remediate through standard fairness remediation techniques for classifier fairness (Hort et al., 2022).

To mitigate this gap, we benchmark the effectiveness of prompting as well as regularization-based remediation methods. In the prompting methods, we rely on instruction-following (Kojima et al., 2022) and study the effectiveness of group agnostic and group aware prompts. We find that these prompt-based methods are unable to decrease the FPR gap - with Muslim and Jewish groups still having FPR about 40% higher than the majority Christian group. Then we study two in-processing (Prost et al., 2019; Beutel et al., 2019) and post-processing (Tifrea et al., 2024) remediation methods by learning a softmax classification layer on the final embeddings of the LLMs, which achieve better fairness-performance trade-offs. Note that the in-processing baseline is only applicable to the fine-tuned model, and cannot be applied in the zero/few-shot classification with no training.

**Contributions.** Our contributions are:

- We benchmark the group fairness of LLM-based

classifiers and show that they do not satisfy Equality of Opportunity along identity aspects such as religion, race, ethnicity, sex.

- We introduce three remediation techniques: prompt-based, in-processing, and post-processing. We highlight that prompting-based remediations which rely on instruction tuning cannot achieve lower false positive rates, and that methods that operate on the embeddings of the LLMs (in/post-processing) are needed to achieve better fairness-performance trade-offs.
- Our findings suggest that prompt-based methods are not effective for group fairness remediation. Additionally, in-processing remediation achieves better fairness-performance trade-offs than post-processing methods.

**Related Work.** Tamkin et al. (2023) provided a method for evaluating how biased a language model may be by generating hypothetical prompts with group information. This was then used to make decisions by fitting a mixed effects model and quantified the degree of bias. 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 mediating zero-shot and few-shot classifiers drawn from LLMs with respect to equality of opportunity fairness. The related work for fairness in classical classification can be found in Appendix A.

## 2 Problem Setup

Throughout this paper, we use PaLM 2 S (Anil et al., 2023) as the base model. We also use PaLM 2 L model in a model transfer experiment. We explore two settings: a ‘zero-shot’ setting, and a ‘fine-tuned’ setting where we improve the classifier performance by training a classifier, or head, on top of the final representation of the model.

For both the zero-shot and fine-tuned case we begin by wrapping the text snippet to be classified in ‘wrapper text’ that encourages the LLM to make a decision. As an example, consider a post with the text ‘*first post!*’ In this toy example, in order to encourage the LLM to make a decision, we wrap the text to obtain ‘is “*first post!*” toxic? please

respond with Yes or No.’ We then send this to an LLM and ask for either a text answer (zero-shot) or model activations (fine-tuning) from the LLM.

**Zero-Shot Classifier.** We use a scoring-based zero-shot method for adapting LLMs to classification tasks. A schematic diagram is given in Fig. 2a in the appendix. Our goal is to take a text post and a classification target as input and return a distribution over negative and positive classes as output. To accomplish this, we send wrapped text to an LLM and ask it to score the ‘Yes’ and ‘No’ tokens to obtain positive and negative class scores. We then apply a softmax to obtain a probability distribution and use that distribution when making decisions. This is similar to the methodology that is followed by (Chung et al., 2022).

**Fine-Tuned Classifier.** In the fine-tuned case, we use the activations of the final layer as input, and train a logistic regression model to predict the ground truth labels. The logistic regression model is then used for decision making. This approach can be thought of as a form of parameter efficient supervised fine tuning. A schematic diagram is given in Fig. 2b in the appendix.

## 3 Experiment Setup

**Dataset and 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.

**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:

$$FPR\ Ratio = \frac{FPR_{group}}{FPR_{majority}}. \quad (1)$$

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.

**Group fairness remediation baselines.** We now evaluate several remediation methods. We propose a family of prompt-based methods that ask the model to be fair at different levels of specificity (in general, to demographic groups, and to a specific group). We also include in-processing and

post-processing methods that are adapted from successful classical approaches:

- *Post-processing.* We apply FRAPPE with an MMD regularizer base (Tifrea et al., 2024).
- *In-processing.* We apply the in-processing MMD regularization of Prost et al. (2019) to model fine-tuning. This baseline is inapplicable to zero/few-shot prompt-based classifiers.
- *Prompt PBF.* ‘Please be as fair as possible when making a decision’ is appended to the prompt.
- *Prompt 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.
- *Prompt PBF2TG.* ‘Please be as fair as possible when making a decision about comments that mention Judaism or Jewish people’ is appended to the prompt.

For more details about the remediation methods, please see Appendix E.

## 4 Experiment Results

**Benchmarking group fairness without remediation.** First, we evaluate PaLM 2 S model (Anil et al., 2023) with respect to equality of opportunity. These experiments follow the classification methodology described in Sec. 2 and use the Civil Comments Identity (Borkan et al., 2019) dataset.

| Group  | Zero-shot FPR Ratio | Fine-tuned FPR Ratio |
|--------|---------------------|----------------------|
| Muslim | 1.89                | 2.24                 |
| Jewish | 1.48                | 1.71                 |

Table 1: False positive rate ratios that quantify the magnitude of violation of equality of opportunity 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 results of this evaluation for the two groups with the highest ratio gaps are given in Tab. 1 and the full table is given in Tab. 2 (in the appendix). These elevated ratios imply that there is headroom for group fairness remediation. In the rest of the experiments, we will describe remediation approaches and present their empirical performance.

**Comparing the Pareto frontiers of group fairness remediation.** Next, we benchmark the different remediation techniques for performance vs group fairness trade-offs. The results of this are

shown in Fig. 1a. For in-processing and post-processing methods, each point in the plot is generated by varying the regularizer strength (see Appendices E.2 and E.3, respectively for more details). We observe that both methods improve fairness without significantly degrading the performance of the classifier.

We also use the activations of the final layer of the LLM to fine-tune a classifier as described in Sec. 2. The in-processed technique remediates during fine-tuning; the post-processed technique performs a separate remediation step after fine-tuning is complete. Note that this fine-tuned model generally performs better than drawing predictions directly from the LLM as described in Sec. 2. However, it requires access to a supervised dataset for fine-tuning. We compare the Pareto frontiers of fairness and performance for each remediation technique in Fig. 1b. As before, each point in the plot is generated by varying the regularizer strength, the term that trades off between performance and fairness terms in each technique’s loss function.

There are a few takeaways from these experiments. First, the ability to fine-tune on LLM-derived activations improves the performance of the classifier as compared to zero-shot case. Second, as before, we are able to improve performance without significantly degrading the performance of the fine-tuned classifier. Third, the in-processing technique generally performs better than the post-processing technique in this environment. Finally, prompt-based methods can offer some fairness and performance benefit but are generally less general and less effective than in-processing and post-processing methods.

### Transfer of remediation to an unseen model.

This experiment has a different setup than those described earlier in this section. Here, we no longer make the assumption that the LLM has ‘introspective’ abilities and thus can not return its activations to use as an embedding. Instead, we use a ‘third party’ model to embed the prefix of our query and use these embeddings for any task where they are required. More specifically, we use the Google News 128-dimensional embedding model (Bengio et al., 2000).<sup>1</sup>

This enables us to operate in environments where drawing embeddings from the LLM is not possible. One interesting case is where we train a post-processing remediation model on one LLM then

<sup>1</sup><https://www.kaggle.com/models/google/nlm>.

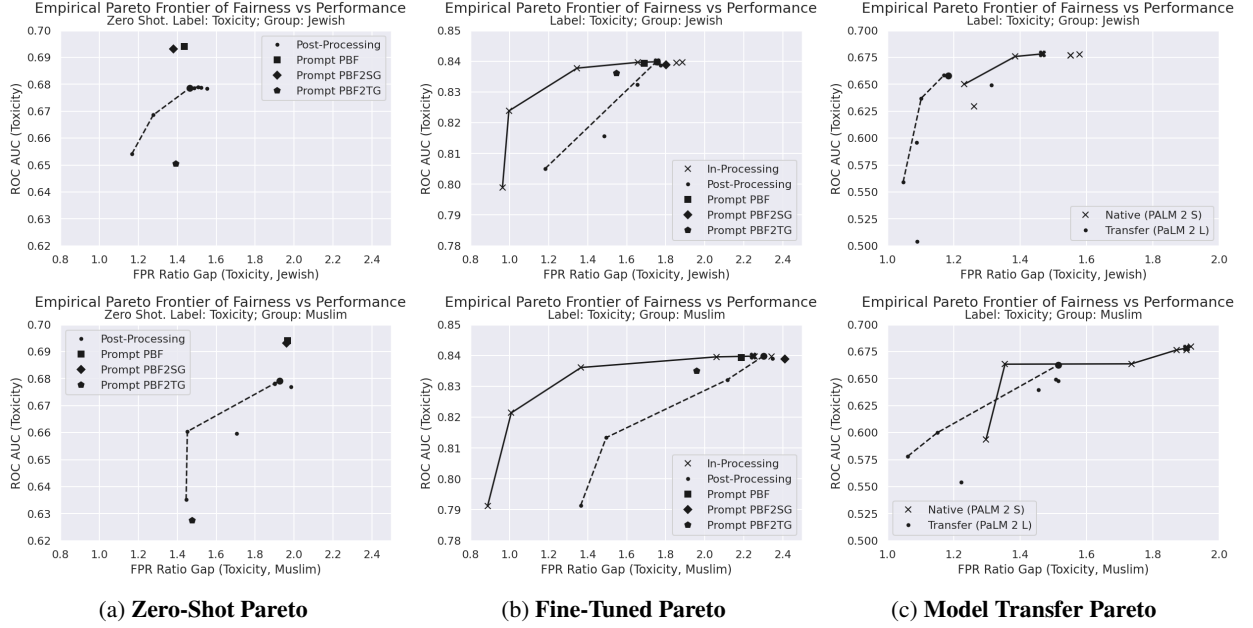


Figure 1: A comparison of the Pareto frontiers of in-processing and post-processing techniques. The left plot shows the performance and fairness of zero-shot remediation and the middle plot shows fine-tuned remediation; top plots give results for the Jewish group and bottom plots gives results for the Muslim group. Prompt-based methods are single points indicated by large symbols. Note that the in-processing baseline is inapplicable in the zero-shot case. Each point for in-processing and post-processing is generated by setting different values for  $\lambda$  in Equations (2) and (4) in the appendix. The bold point gives the unremediated ( $\lambda = 0$ ) case for in-processing and post-processing. 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).

apply to another; can we reuse the existing fairness model to improve fairness with the new model? Fig. 1c gives the results this model transfer learning scenario. 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 can save on inference costs by training our fairness model on a LLM with fewer parameters then apply it to a larger LLM at inference time.

Note also that the Pareto frontier achieved by the smaller PaLM 2 S model in Fig. 1c, where ‘third party’ embeddings are used, is just slightly degraded from the Pareto frontier given in the upper plot of Fig. 1a where model activations are used as embeddings. This suggests that using ‘third party’ embeddings is a reasonable choice for remediation if model activations are not available.

## 5 Concluding Remarks

We study the fairness of LLM-based classifiers. We identify that LLM-based classifiers may exhibit group unfairness. We introduce three remediation

techniques to improve fairness while maintaining acceptable performance for LLM-based classifiers. We find that prompt-based techniques offer limited benefit and are outperformed by in-processing and post-processing techniques. We also find that the in-processing technique consistently provides favorable performance vs fairness trade-off in the fine-tuned settings.

Based on our results, if remediating a fine-tuned model, in-processing seems to be a more robust approach because it consistently provides favorable fairness versus performance tradeoffs. In other LLM-based classification settings where in-processing cannot be applied (such as zero-shot remediation and transfer tasks) the post-processing technique seems to be a promising approach.

We generally 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 very sur-



prising given that fairness is a distributional issue, and hence simple prompting may not necessarily provide the distribution matching effects that we expect from remediation.

## 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 prompt-based 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 prompt-based methods that work. Furthermore, it has been observed that LLM capabilities improve as LLM size increases (Wei et al., 2022b), and this could have a beneficial effect on prompt-based method effectiveness as LLMs become larger and more capable.
- Our experiments are focused on equality of opportunity (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.’
- LLM-based classifier inference is very expensive when compared to simpler models, and the performance of LLM-based classifiers does not yet justify that cost (for example, our LLM-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 LLM-based classifiers will improve enough over time to justify their high inference costs and become deployed systems where fairness considerations are in play.
- We consider only one LLM (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 case, the need for developing fairness remediation techniques 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 compare against 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.

## References

- Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426.
- Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. 2022. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760.
- 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 Crepey, 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. 2023. [Palm 2 technical report](#). *Preprint*, arXiv:2305.10403.
- 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. 2022. [Constitutional ai: Harmlessness from ai feedback](#). *Preprint*, arXiv:2212.08073.
- Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. 2021. Your fairness may vary: Pretrained language model fairness in toxic text classification. *arXiv preprint arXiv:2108.01250*.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.
- Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.
- Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459.
- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.
- Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001.
- Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. 2020. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Evgenii Chzhen and Nicolas Schreuder. 2020. A min-max framework for quantifying risk-fairness trade-off in regression. *arXiv preprint arXiv:2007.14265*.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. [Toxicity in chatgpt: Analyzing persona-assigned language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.
- Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801.
- Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. [Making pre-trained language models better few-shot learners](#). In *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)*, pages 3816–3830, Online. Association for Computational Linguistics.
- Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. 2020. Learning unbiased representations via Rényi minimization. *arXiv preprint arXiv:2009.03183*.
- Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. 2019. Fairness-aware neural Rényi minimization for continuous features. *arXiv preprint arXiv:1911.04929*.
- Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2022. Bias mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR.

- Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. [Improving diversity of demographic representation in large language models via collective-critiques and self-voting](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, Singapore. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohata, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.
- Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. 2022. A stochastic optimization framework for fair risk minimization. *Transactions of Machine Learning Research*.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *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)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.
- Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. 2019. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*.
- Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems*, 30.
- Edward Raff, Jared Sylvester, and Steven Mills. 2018. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 243–250.
- Goce Ristanoski, Wei Liu, and James Bailey. 2013. Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1529–1532.
- 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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.
- Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. 2020. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. [Finetuned language models are zero-shot learners](#). *Preprint*, arXiv:2109.01652.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. 2023. [Bloomberggpt: A large language model for finance](#). *Preprint*, arXiv:2303.17564.

- Ruicheng Xian, Lang Yin, and Han Zhao. 2023. Fair and optimal classification via post-processing. In *International Conference on Machine Learning*, pages 37977–38012. PMLR.
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.
- Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.
- Alexandru Tifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. 2024. [FRAPPÉ: A group fairness framework for post-processing everything](#). *International Conference on Machine Learning*.



## A Related Work in Classical Classification Fairness

There is rich literature on fairness in ‘classical’ (i.e. non-LLM-based) classification. In this paper we focus on the ‘equality of opportunity’ notion of group fairness (Hardt et al., 2016) 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 and 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 and 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; Tifrea 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 LLM-based classifiers. Having said that, most post-processing approaches require access to demographic labels at test time, which is infeasible, especially for LLM-based classifiers. That is why we only focused on FRAPPÉ (Tifrea et al., 2024) which works without such access.

## B 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.

## C LLM-Based Classification Methodology

Fig. 2 gives a schematic diagram of the classification methodology described in Sec. 2. Note that we use the following wrapper text for our toxicity classifier: “‘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.”

## 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 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.

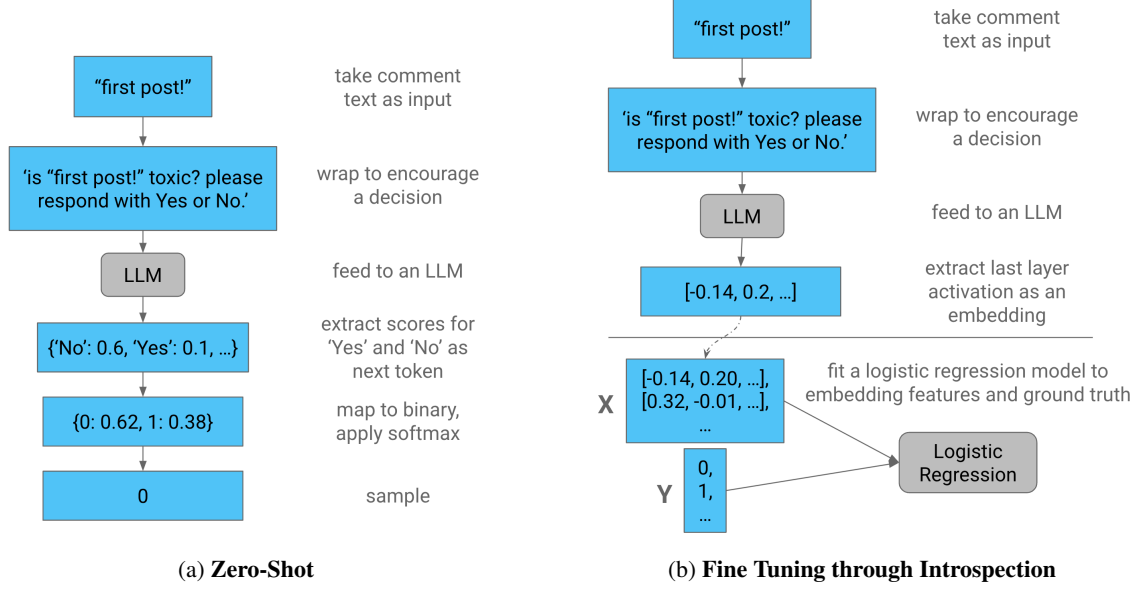


Figure 2: Classification flow diagrams for zero-shot and fine-tuned LLM-based classification. In both cases, decisions are encouraged via ‘text wrappers’ that nudge the LLM to make a classification decision. In the zero-shot case (Fig. 2a) we treat the wrapped text as a prefix and query the LLM for two 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. In the fine-tuned case (Fig. 2b) we assume that the LLM is ‘introspective’ and can supply its activations. We instead query the LLM 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.

## E 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.

### E.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.

### E.2 In-Processing

Min Diff (Prost et al., 2019) is an approach that has been successful in remediating to achieve equality of opportunity in the ‘classical’ setting. 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 in 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 the equations that follow, let  $(x, y)$  represent a feature and a label, where  $x \in \mathcal{X}$ , and  $y \in \{0, 1\}$ .

We use the following loss:

$$L_{IP} = L_{CE}(\hat{Y}, Y) + \lambda MMD \left( \hat{Y}|Y = 0, G = 0; \hat{Y}|Y = 0, G = 1 \right). \quad (2)$$

Where  $L_{IP}$  is the in-processing loss,  $L_{CE}$  is the usual cross entropy loss,  $MMD$  is a Maximum Mean Discrepancy kernel that gives the distance between the distributions of probability of a false positive for group and associated majority, and  $\lambda$  is a parameter that trades off between the two loss terms (and thus between performance and fairness).

In order to adapt this to the LLM case, we use the Min Diff loss during fine tuning. This approach does not work in the zero-shot case where no fine tuning is performed, but the following approach does.

### E.3 Post-Processing

Recent work has demonstrated how in-processing techniques can be adapted to post-processing scenarios (Tifrea 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:

$$P_{pp}(y|x) = \sigma(z(P_{bl}(y|x)) + z(P_{ef}(y|x))), \quad (3)$$

where  $\sigma$  is the logistic function,  $z$  is the logit function,  $P_{bl}(y|x)$  is the baseline prediction distribution,  $P_{ef}(y|x)$  is the emfairening model’s distribution, and  $P_{pp}(y|x)$  gives the combined post-processed distribution. The emfairening model can be trained with the following loss:

$$L_{PP} = KL(P_{bl}(y|x)||P_{pp}(y|x)) + \lambda MMD \left( \hat{Y}|Y = 0, G = 0; \hat{Y}|Y = 0, G = 1 \right), \quad (4)$$

where  $KL$  is the Kullback-Leibler divergence and  $MMD$  is the Maximum Mean Discrepancy kernel. As with the Min Diff approach, the Maximum Mean Discrepancy kernel pushes the probabilities of a false positive for group and majority (and thus false positives rates for group and majority) toward each other. 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,  $\lambda$  trades off between the two loss terms and thus dials between fairness and performance.

Note that we are free to choose the baseline model  $P_{bl}(y|x)$ . As such, we can apply this approach directly to the LLM in a zero-shot environment or apply it to a fine-tuned model.

## F 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. 4. The full table is given in Table 2

## G 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.

| <b>Group</b>              | <b>Zero-Shot FPR Ratio</b> | <b>Fine-Tuned FPR Ratio</b> |
|---------------------------|----------------------------|-----------------------------|
| 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 a classifier built on PaLM 2.