FairAdapt: An Exploration of Fairness-promoting Adapters

Anonymous ACL submission

Abstract

Large language models work better for some than others, and lightweight mitigation of performance disparities across social groups could help bridge inequality gaps. Here, we explore fairness-promoting adapters as a potential mitigation technique. We find that generally adapters lead to as good or better performance than full fine-tuning, with mixed effects on group disparity. Combining fairness-promoting adapters does *not* lead to smaller group disparity, and while Adapter Fusion is superior to model stipulation, such systems fail to outperform non-fairness promoting adapters. Combinations of fairness-promoting adapters seem to positively effect group fairness under temporal concept drift, although, as expected, we observe a generalized performance drop. From the perspective of group fairness, our results are somewhat negative, and we discuss the potential bottlenecks for current approaches to mitigating group disparity.

1 Introduction

The roll-out of language models in recent years has raised concerns around fairness and equity, particularly across societal groups defined by protected attributes such as gender and race. The imperative to ensure fairness, *i.e.*, equal performance, across such groups has gained substantial traction. On the other hand, language models have grown in size and the cost of mitigating biases and correcting for performance disparities has increased. The need for efficient, light-weight mitigation strategies is plain to see.

Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020, 2021) have been a prominent technique to efficiently fine-tune transformer-based language models. Adapters are trained to solve target tasks they are fine-tuned on, on top of fixed representations from existing models. Only a small set of new parameters are introduced –usually less than 2% of the total number of model parameters.

Previous work (Lauscher et al., 2021; Kumar et al., 2023; Hauzenberger et al., 2023) has investigated on-demand modular debiasing methods, achieving on-par task performance compared to the non-debiased models. In this work, we investigate whether adapters can be used to mitigate performance disparities across societal groups. To this end, we introduce adapters trained with fairness-promoting objectives such as, for example, group distributionally robust optimization or spectral decoupling. We evaluate such adapters and combinations thereof on two legal classification datasets from the FairLex benchmark (Chalkidis et al., 2022), namely crime severity prediction in Chinese, and legal outcome forecasting in German. In doing so, we aim to answer the following research questions:

R1: *Do fairness-promoting adapters improve performance parity?* We train adapters with empirical risk minimization and four fairness-promoting objectives across two datasets, comparing adapters to full fine-tuning.

R2: *Do combining fairness-promoting adapters through Adapter Fusion (Pfeiffer et al., 2021) or model stipulation improve performance parity?* We combine fine-tuned adapters, originally trained individually with different fairness-promoting algorithms, and assess their performance concerning fairness compared to the individual ones.

R3: *What is the effect of the temporal concept drift in empirical fairness?* Since our datasets are chronologically split into training, validation, and test sets, we consider the potential performance decrease in the latter ones (validation, test) and the fluctuation in group disparities. We should of course expect a small drop from validation to test, but we will analyze relative differences in drop sizes as the effect of temporal concept drift.[1]

[1]Theoretically, relative differences could also be caused by differences in proneness to overfitting.

Contributions and Findings We evaluate a promising set of techniques for mitigating performance disparities across social groups in pre-trained models: adapters with fairness-promoting objectives and combinations thereof. We find that only one of the fairness-promoting adapters consistently reduce performance disparities across-social groups: adapters trained with group distributionally robust optimization. Spectral decoupling sometimes leads to increased fairness, but not robustly. Combinations of different objectives did not seem effective in mitigating disparities. As for sensitivity to temporal concept drift, we found that performance disparities did not increase over time but rather decreased. In sum, we find some support for **R1** (for G-DRO) and **R3** (group disparity), and *no* support for **R2**.

2 Experiments

Standard language models are trained to minimize the average training loss via empirical risk minimization (ERM), *i.e.*, vanilla cross-entropy. Many other learning algorithms have been proposed to overcome one of the main ERM's shortcomings: ERM is prone to overfitting to spurious correlations and, therefore, unable to generalize well across domains or subpopulation shifts. In this work, we explore the use and combination of adapters optimized with standard fairness-promoting algorithms.

Adapters We base the implementation of adapters on work from Pfeiffer et al. (2021), as well as their Adapter Fusion strategy to combine them.

Fairness NLP researchers have uniformly adopted John Rawls' definition of fairness (Williamson and Menon, 2019; Ethayarajh and Jurafsky, 2020; Cabello and Søgaard, 2022; Chalkidis et al., 2022), defining fairness as performance parity, except where it worsens the conditions of the least advantaged. We do the same and evaluate group fairness quantifying performance differences across demographic groups, referred to as group disparities, while also looking at worst-group performance (measured as macro-F1).

Datasets We experiment with two classification datasets, which are part of FairLex (Chalkidis et al., 2022), a benchmark for the evaluation of empirical fairness on legal NLP tasks. CAIL, originally published by Wang et al. (2021b), comprises approx. 100k criminal cases from China. The task is *crime severity* prediction, a multi-class classification task, where given the facts of a case, the goal is to predict how severe the committed crime is from 0 to 5. We examine fairness with respect to two demographic attributes: (a) the *region* of the court, and (b) the *gender* of the defendant. FSCS, originally published by Niklaus et al. (2021), comprises approx. 80k cases from the Federal Supreme Court of Switzerland (FSCS). The task is *legal judgment forecasting*, in which case is a binary classification task considering the *approval*, or *dismissal* of a case (appeal). We consider the subset of cases written in German, approx. 35k, and examine fairness with respect to two demographic attributes: (a) the *region* of the court, and (b) the *legal area* relevant to the case.

Pre-trained Models We use the domain-specific transformer-based language models released by Chalkidis et al. (2022). Chalkidis et al. released individual MiniLM (Wang et al., 2021a), distilled versions of XLM-R (Conneau et al., 2020), which have been further pre-trained in domain-specific corpora, *e.g.*, Chinese criminal cases for CAIL. We use these models as our baselines and either fully fine-tune them, or fine-tune plug-in adapters.

Fairness-promoting Algorithms We consider four alternative fairness-promoting algorithms that are either attribute-aware, *i.e.*, demographic annotations are needed and used, or attribute-unaware, *i.e.*, demographic annotations are not needed and not used.

i) *Attribute-aware methods:* Group distributionally robust optimization **G-DRO** (Sagawa et al., 2020) accounts for group-wise losses using adaptive group weight. We couple models based on G-DRO with strong L2 regularization to improve worst-group generalization, as suggested in Sagawa et al. (2020). Invariant risk minimization **IRM** (Arjovsky et al., 2020) learns a feature representation such that the optimal classifier, on top of that feature representation, matches across data distributions.

ii) *Attribute-unaware methods:* Spectral Decoupling **SD** (Pezeshki et al., 2020) introduces a regularization loss term that helps to mitigate garadient starvation, a phenomenon that emerges when training with cross-entropy loss. Risk Extrapolation **REx** (Krueger et al., 2020) introduces a penalty on the variance of training risks to make the model more robust to distributional shifts.

	CAIL (ZH)						FSCS (DE)						
ALGORITHM	REGION			GENDER			REGION			LEGAL A.			
	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{GD}}$
FULL FINE-TUNING													
ERM	61.4	**57.3**	**3.9**	61.4	**60.7**	0.6	67.8	58.7	4.4	67.8	55.1	8.7	4.4
G-DRO	60.5	55.9	4.1	**61.9**	60.3	0.8	67.8	63.7	3.1	64.0	**56.2**	8.5	4.1
IRM	61.0	54.3	5.2	61.0	59.3	0.9	66.3	**59.0**	4.9	63.0	51.8	**7.2**	4.6
SD	61.4	55.3	5.1	61.6	60.6	**0.5**	67.9	63.6	**2.1**	67.9	54.6	8.9	4.1
REx	**61.6**	55.3	4.5	61.5	59.9	0.8	**68.0**	61.4	3.3	**68.0**	55.3	8.8	4.4
ADAPTERS													
ERM	61.1	49.9	6.2	**61.1**	**61.2**	**0.2**	67.7	63.9	3.5	67.7	52.0	10.1	5.0
G-DRO	59.6	54.4	**3.2**	56.2	54.0	1.0	67.3	62.6	3.5	60.6	53.2	**4.3**	**3.0**
IRM	60.1	52.0	5.8	56.1	51.1	2.7	65.9	62.0	**2.4**	66.2	52.1	9.7	5.2
SD	55.8	48.8	4.8	56.1	53.1	1.5	**69.5**	**65.4**	2.9	**69.5**	**54.6**	10.0	4.8
REx	**62.1**	**56.1**	4.5	57.2	51.7	3.0	68.1	62.5	3.5	68.1	51.7	10.1	5.3
ENSEMBLE OF PAIRS OF FAIRNESS-PROMOTING ADAPTERS													
IRM,G-DRO	60.5	55.5	4.7	62.0	61.8	0.3	66.9	61.8	3.6	62.1	54.5	8.2	4.2
SD,REx	61.0	53.2	5.2	**62.6**	**62.4**	0.2	62.6	46.5	8.2	64.4	55.4	9.0	5.7
FUSION OF FAIRNESS-PROMOTING ADAPTERS													
IRM,G-DRO	60.2	55.9	**3.4**	61.8	61.7	**0.1**	67.2	60.6	4.1	61.7	52.4	10.2	4.5
SD,REx	**62.1**	56.1	4.8	58.1	54.0	2.2	68.5	61.8	4.1	68.5	54.5	10.3	5.4
ENSEMBLE OF TOP-3 ADAPTERS													
ERM,G-DRO,SD	60.0	55.5	4.7	61.0	60.1	0.8	67.0	62.6	3.8	64.4	54.5	10.1	4.9

Table 1: Validation results for all learning algorithms per dataset attribute. We report the average performance across groups ($\overline{\text{mF1}}$), the worst-group performance (mF1_w) and group disparity among groups (GD). Best overall metric is in **bold**; **best** and second-best metrics within each tuning strategy (FULL or ADAPTERS) are also marked. In FULL FINE-TUNING, only 2/4 fairness-promoting objectives (G-DRO and SD) reduce group disparity on average. G-DRO and SD are also the best adapters; with G-DRO being by far most fair. None of the ensembles, including Adapter Fusion, succeed in lowering group disparity.

Combination strategies We evaluate three different strategies for combining individually trained adapters: (a) a post-hoc ensemble of pairs of adapters, where we aggregate their output (soft) probabilities before making a prediction, (b) Adapter Fusion, as presented in Pfeiffer et al. (2021), and (c) a post-hoc ensemble of the three best-performing algorithms, including ERM, where the label prediction is based on a majority vote.

3 Results

In Table 1, we present validation results across all datasets, attributes, and learning algorithms. Table 2 in Appendix A shows results on the test sets.

Full Fine-tuning Focusing on the results for all learning algorithms (top group of Table 1) in the full fine-tuning setting, we observe that results are mixed, and the application of fairness-promoting algorithms do not always improve empirical fairness, if not the opposite, compared to ERM. In general, worst-group performance (mF1_w) and group disparity (GD) improve in 6 out of 16 cases –note that this improvement mainly happens on FSCS data–. This is in line with the literature (Chalkidis et al., 2022; Brandl et al., 2023), where they also find that in many cases, when these algorithms that intend to improve fairness are applied to realistic datasets, fail to do so.

Adapter-based Fine-tuning Comparing the results of full fine-tuning with ERM (top group of Table 1) to those of adapter-based fine-tuning with ERM (second top group of Table 1), we observe that adapter-based fine-tuning improves the worst-group performance 2 out of 4 times without

severely hurting the overall performance ($\overline{\text{mF1}}$). GD is also reduced in the same two cases.

Moving to the application of fairness-promoting algorithms to adapter-based fine-tuning, we observe that worst-group performance (mF1_w) only improves in 7 out of 16 cases. Group disparity is consistently reduced (or remains the same) when considering court region in both datasets, and legal area in FSCS. Specifically, attribute-aware fairness-promoting algorithms (G-DRO, IRM) succeed on reducing group disparities further than attribute-unaware algorithms (SD, REx). While the latter perform better in terms of worst-group performance, which is not surprising since the overall performance increases as well.

Results for gender groups in CAIL point to a general negative effect of fairness-promoting algorithms when targeting binary groups.

Combining Adapters Combination of individual adapters is done following different strategies, and therefore we expect different behavioral outcomes. In the three lower parts of Table 1, we observe that Adapter Fusion is beneficial in terms of average performance ($\overline{\text{mF1}}$), but it generally yields lower mF1_w compared to the naive ensembles. As for group disparity, results are mixed. We speculate that the general worsen in empirical fairness could be due to the *knowledge composition* step (Pfeiffer et al., 2021) based on ERM, which is performed on top of the optimized fairness-promoting adapters.

The effect of the temporal concept drift Training, validation and test splits are chronologically split; these chronological splits entail a label distribution shift for a given group. While we confirm an overall performance decrease with the corresponding lower worst-group F1 scores, we also find that group disparities are reduced. In other words, the effect of temporal concept drift smoothen the differences in performance across groups. Figure 1 provides a visual for a comparison. Combinations of fairness-promoting adapters seem to negatively effect group fairness under temporal concept drift.

4 Discussion: What Goes Wrong?

G-DRO adapter mitigation showed moderate success in our experiments, but most fairness-promoting objectives (and combinations thereof) failed to reduce performance gaps between groups. Why is that? Here, we list some of the options.

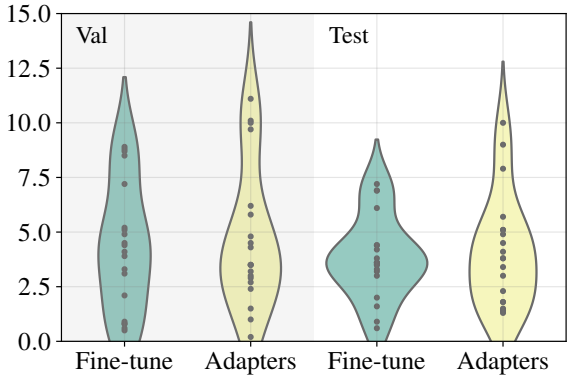


Figure 1: Variance in group disparity. We can observe the effect of temporal drift by comparing results between validation (left) and test (right).

Algorithmic limitations Fairness-promoting learning algorithms have only been around for a few years, and it is conceivable that the right objectives for the sort of problems considered here, have not been found yet.

Dataset limitations Our experiments make two assumptions: a) that social group membership correlate with language and legal outcome, and b) that we have enough data to learn these correlations. We know there are correlations between group membership and model performance, but they may be subtle and hard to model robustly. It is therefore very likely that more data may be needed to learn the relevant patterns.

Do protected attributes cut at the joints? Much work on fairness assumes variation is along the axis of protected attributes such as gender and race, but of course, this may not be true. Perhaps variation is primarily along other dimensions such as literacy level or professional interest.

5 Conclusion

This paper presented a comparative analysis of adapters optimized with standard fairness-promoting algorithms. We explored the use and combination of adapters, and how their empirical fairness compare to full fine-tuning a model. Our findings suggest that attribute-aware algorithms, such as G-DRO, are the most viable approach to mitigate group disparities whenever group membership information is available. However, there is a need for more effective light-weight strategies to reliably mitigate biases and group disparities.

Limitations

The study presented in this paper is general and extensible to analyse other forms of performance inequalities in language models. We root the experiments in one model architecture (Wang et al., 2021a) and one adapter (Pfeiffer et al., 2021). However, our work would benefit from analyzing a wider range of models and parameter-efficient training strategies.

We consider two datasets from the legal domain, with well defined demographic attributes. Further research on the interaction between parameter-efficient training and fairness-promoting algorithms should account for the application to other domains, where the conceptualization of fairness might differ.

Additional experiments would help to gain a better insight. For instance, accounting for the variance of the fine-tuning processes –for both full fine-tuning and addapters– when varying the random weight initialization.

Ethics Statement

The models and datasets used in this study are publicly available, and we strictly follow the ethical implications of previous research related to the data sources. We do not anticipate other ethical risks derived from our work.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization.

Stephanie Brandl, Emanuele Bugliarello, and Ilias Chalkidis. 2023. On the interplay between fairness and explainability.

Laura Cabello and Anders Søgaard. 2022. Are pre-trained multilingual models equally fair across languages? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3597–3605, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. 2023. Modular and on-demand bias mitigation with attribute-removal subnetworks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6192–6214, Toronto, Canada. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Rémi Le Priol, and Aaron C. Courville. 2020. Out-of-Distribution Generalization via Risk Extrapolation (REx). *CoRR*.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Pezeshki, Sekouba Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. 2020. Gradient starvation: A learning proclivity in neural networks. In *Neural Information Processing Systems*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021a. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021b. Equality before the law: Legal judgment consistency analysis for fairness. *Science China - Information Sciences*.

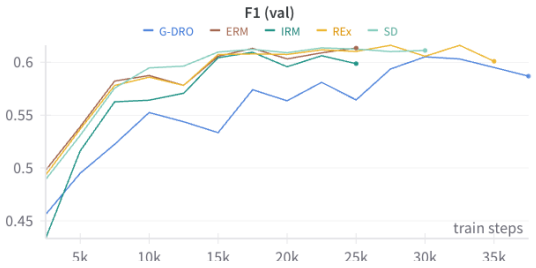
Robert Williamson and Aditya Menon. 2019. Fairness risk measures. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797, Long Beach, California. PMLR.

A Test Performance

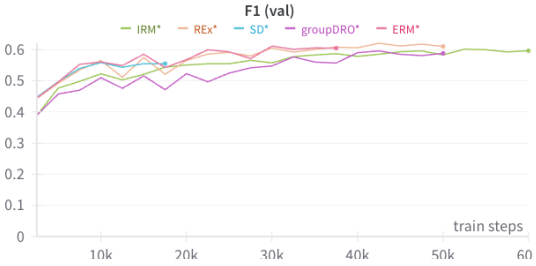
Table 1 shows test results across all datasets, attributes, and learning algorithms.

B Learning curves

In Figure 2 we represent the evolution of learning curves (F1 scores on the validation set) during training. While adapters need more training steps to converge, the ceiling effect is also more prominent compare to full fine-tuning the model. G-DRO, IRM, and REx have a tendency to overfit, specially in full fine-tuning.



(a) Full fine-tuning



(b) Adapters

Figure 2: Evolution of macro-F1 scores, evaluated on the validation split, when fine-tuning on CAIL. Fairness-promoting algorithms target *court region* groups.

	CAIL (ZH)						FSCS (DE)						
ALGORITHM	REGION			GENDER			REGION			LEGAL A.			
	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{mF1}}$	mF1_w	GD	$\overline{\text{GD}}$
FULL FINE-TUNING													
ERM	58.9	54.8	3.6	58.9	58.1	3.5	63.7	58.8	3.2	63.7	51.3	7.2	4.4
G-DRO	58.4	50.5	4.2	57.4	56.4	**0.6**	**64.5**	58.3	3.3	61.4	**52.0**	6.9	3.8
IRM	58.8	53.2	4.4	59.8	59.3	1.6	62.4	57.8	**2.0**	59.9	51.0	**6.1**	3.5
SD	**59.7**	**56.9**	**3.0**	**60.0**	**59.8**	0.9	63.2	58.8	3.6	63.2	51.2	6.9	3.6
REx	59.6	56.4	3.3	59.7	59.0	2.6	63.8	**61.2**	2.8	**63.8**	51.3	7.4	4.0
ADAPTERS													
ERM	58.7	53.1	5.7	**59.3**	**58.9**	1.4	65.1	58.0	3.4	65.1	49.0	10.0	5.1
G-DRO	57.7	52.8	4.1	52.7	50.4	1.3	64.9	**62.8**	**1.4**	59.0	**51.5**	**5.1**	3.0
IRM	58.4	53.5	**3.8**	52.7	52.3	1.5	63.6	60.2	1.8	63.7	50.9	7.9	3.8
SD	52.5	44.4	4.9	52.5	52.0	1.8	64.8	60.9	2.3	64.8	50.8	9.0	4.5
REx	**60.2**	**54.4**	4.4	54.2	53.9	**0.9**	**66.3**	61.5	2.6	**66.3**	50.7	10.0	4.5
ENSEMBLE OF FAIRNESS-PROMOTING ADAPTERS													
IRM,G-DRO	60.1	55.8	4.8	59.1	57.9	1.2	63.7	59.3	2.7	60.5	**53.1**	5.5	3.5
SD,REx	60.3	56.8	3.1	60.0	59.0	1.0	63.4	59.2	2.7	61.1	49.5	8.0	3.7
FUSION OF FAIRNESS-PROMOTING ADAPTERS													
IRM,G-DRO	59.6	54.1	3.5	58.8	58.0	**0.4**	63.2	59.3	2.7	59.5	52.8	7.2	3.5
SD,REx	**61.3**	55.7	5.2	56.5	55.5	0.6	63.5	59.3	2.7	63.5	48.9	9.3	4.5
ENSEMBLE OF TOP-3 ADAPTERS													
ERM,G-DRO,SD	58.2	53.2	4.3	**60.2**	58.5	1.7	64.6	61.0	2.3	61.0	50.6	8.8	**2.8**

Table 2: Test results for all learning algorithms per dataset attribute. We report the average performance across groups ($\overline{\text{mF1}}$), the worst-group performance (mF1_w) and group disparity among groups (GD). Best overall metric is in **bold**; **best** and second-best metrics within each tuning strategy (FULL or ADAPTERS) are also marked.