On Mitigating Performance Disparities in Multilingual Speech Recognition

Monorama Swain, Anna Katrine van Zee, and Anders Søgaard
Dpt. of Computer Science
University of Copenhagen
Lyngbyvej 2, DK-2300 Copenhagen
soegaard@di.ku.dk

Abstract

How far have we come in mitigating performance disparities across genders in multilingual speech recognition? We compare the impact on gender disparity of different fine-tuning algorithms for automated speech recognition across model sizes, languages and gender. We look at both performance-focused and fairness-promoting algorithms. Across languages, we see slightly better performance for female speakers for larger models regardless of the fine-tuning algorithm. The best trade-off between performance and parity is found using adapter fusion. Fairness-promoting fine-tuning algorithms (Group-DRO and Spectral Decoupling) hurt performance compared to adapter fusion with only slightly better performance parity. LoRA increases disparities slightly. Fairness-mitigating fine-tuning techniques led to slightly higher variance in performance across languages, with the exception of adapter fusion.

1 Introduction

Automatic Speech Recognition (ASR) systems have become ubiquitous in our daily lives, powering virtual assistants, dictation software, customer service interactions, and more. However, these systems are not always equally effective for all users, and gender disparity in their performance is a significant concern (Tatman, 2017).

One key factor contributing to gender disparity in ASR performance is the way in which these systems are trained. If the training data predominantly consist of male voices, for example, ASR systems may exhibit higher error rates when transcribing female speech. Or vice versa. The downstream societal impact of gender disparity in ASR systems is multifaceted. It can exacerbate existing inequalities by hindering access to information and services for certain social groups. In professional settings, inaccurate transcription can impede communication and productivity, potentially affecting career advancement opportunities. Finally, the perpetuation of performance disparities in technology can reinforce harmful stereotypes and norms, contributing to broader social inequalities.

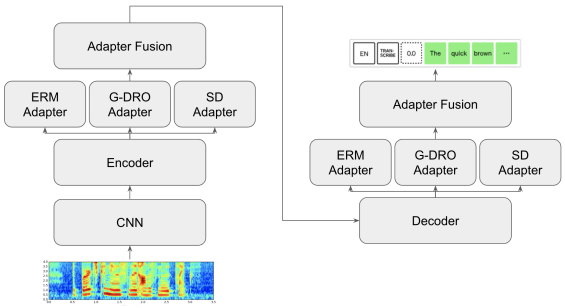

Figure 1: Augmenting Whisper with adapter fusion for better performance and gender parity. Adapter fusion is over three adapters – one trained with a vanilla loss (ERM), one trained with Group-DRO, and one trained with spectral decoupling (SD).

Mitigating gender disparities is non-trivial. Several algorithms have been proposed (Koh et al., 2021), but they often introduce significant performance costs (Kim et al., 2020), and their impact on gender disparity is often small (Chalkidis et al., 2022). In this paper, we evaluate their impact on performance and gender disparity in multilingual ASR, but we also present a simple technique to improve the robustness of fairness-improving fine-tuning methods. Specifically, we augment multilingual Whisper using adapter fusion (Pfeiffer et al., 2021) over several fairness-promoting strategies; see Figure 1.

Contributions We evaluate the performance and gender disparity of four different fine-tuning strategies for multilingual ASR models, two of which are fairness-promoting. Our baselines are standard empirical risk minimization (ERM) and so-called *low-rank adaptation* (LoRA) (Hu et al., 2022). The two fairness-promoting algorithms are Group-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5647–5655

DRO (Sagawa et al., 2020) and spectral decoupling (Pezeshki et al., 2021). Finally, we evaluate a novel combination of existing techniques: adapter fusion over ERM, Group-DRO, and spectral decoupling. We evaluate all models across 16 languages with binary gender as our demographic variable, and find that only adapter fusion reduces gender disparity without incurring a significant performance drop. In fact, performance improves, *and* disparity is reduced. Cross-lingual performance gaps widen a bit, however.

2 Fairness-Promoting Finetuning

Standard expected risk minimization (ERM) fine-tuning in automatic speech recognition (ASR) and fairness-promoting fine-tuning techniques like Group-DRO (Sagawa et al., 2020), spectral decoupling (Pezeshki et al., 2021), and adapter fusion (Pfeiffer et al., 2021) serve different purposes and address different aspects of model performance. ERM fine-tuning, our baseline technique, minimizes the expected risk or the expected value of a loss function over the data distribution. In ASR, this typically involves reducing the word error rate (WER). ERM fine-tuning involves iteratively updating the parameters of the ASR model using a gradient descent-based optimization algorithm, such as stochastic gradient descent (SGD) or Adam, to minimize the loss function.

Fairness-promoting fine-tuning techniques aim to increase performance parity in ASR systems by addressing disparities in model predictions across different demographic variables (e.g., gender, race, age). Group-DRO (Sagawa et al., 2020) modifies the standard fine-tuning objective to optimize for performance parity by minimizing the worst-case error across multiple demographic groups, ensuring that no particular group is disproportionately affected by the model's predictions. Spectral decoupling (Pezeshki et al., 2021) involves modifying the training process to remove unwanted biases from the learned representations, by decoupling sensitive features from the predictive features in the model.

The fairness-promoting techniques differ from ERM fine-tuning in their explicit attention to performance parity, often introducing additional constraints or modifications to the training process to achieve more equitable model outcomes. While ERM fine-tuning seeks to optimize the overall performance of the ASR system, fairness-promoting techniques specifically target performance disparities across social groups.

Adapter fusion fine-tunes a pre-trained model by adding small neural adapters to the model architecture, which are specifically trained to mitigate performance disparities and improve empirical fairness in the model's predictions while retaining the knowledge learned from the pre-training phase. The adapter fusion model explored here combines ERM, Group-DRO and spectral decoupling adapters by adding an extra adapter layer.

3 Experiments

Model Whisper (Radford et al., 2022) is a family of ASR models developed by OpenAI. The models vary in size from 39M parameters in the 'tiny' model to 1.55B parameters in the 'large' model, and they are all trained on 680,000 hours of web data across 97 languages. We focus on Whisper-large because it achieves the highest performance for all groups, but we provide results for all model sizes in Table 3 of the Appendix.

Data We use VoxPopuli[1] – a multilingual, open source dataset – to evaluate the performance and gender disparity of different fine-tuning and fairness-promoting strategies. VoxPopuli is a collection of speeches given in the European Parliament between 2009–2020, and metadata about the speakers demographic information is included. Our study of gender disparity is limited by the availability of only the binary genders in this dataset. We use data from 16 languages: English, Slovene, Lithuanian, Italian, French, Polish, Romanian, German, Finnish, Dutch, Croatian, Hungarian, Slovak, Czech, Spanish, and Estonian. The Whisper models performed relatively poorly out of the box on this data, so we report word error rates for fine-tuned models only.

LoRA and ERM baselines We use two different baselines. Our LoRA baseline introduce low-rank parameterization to the adapter layers added on top of the pre-trained model. Low-rank parameterization reduces the number of parameters required to adapt the pre-trained model, making it more computationally efficient, yet aims to maintain the expressiveness of the adapted model. By constraining the parameters to be low-rank, LoRA like sparsity-promoting regularizers encourages the

[1] https://huggingface.co/datasets/facebook/voxpopuli

adapter layers to capture essential task-specific information. Our ERM baseline adapter layers train additional layers on top of the pre-trained Whisper model. The purpose of these layers is to adapt the pre-trained model's representations to better suit the VoxPopuli data and increase performance. ERM focuses on minimizing the empirical risk, which is the average loss over the training dataset.

Group-DRO and Spectral Decoupling Our two fairness-promoting fine-tuning strategies are also implemented as adapter layers. Unlike the baseline adapters, these strategies specifically target and mitigate group disparity, in our case, across genders. All adapter strategies use the same number of parameters; see Table 2 for the Appendix for hyper-parameters.

Adapter fusion Inspired by the adapter fusion framework presented in Pfeiffer et al. (2021), we introduce adapter fusion layers to adjudicate between three adapters: one trained with empirical risk minimization, one trained with group-distributionally robust optimization, and one trained with spectral decoupling.

Protocol We rely on the standard VoxPopuli splits. For each fine-tuning strategy, we fine-tune a new model for each language. All hyper-parameters are optimized on held-out (development) data, by doing grid search over a limited set of values; see Table 2 of the Appendix for hyperparameters.

Adapter	♀	♂	♀+♂	Δ
LoRA	12.4	13.3	12.9	0.9
ERM	9.9	10.5	10.3	0.6
Group-DRO	10.3	10.5	10.4	**0.2**
SD	10.4	10.8	10.6	0.4
Fusion	**9.4**	**10.0**	**9.7**	0.6

Table 1: Word Error Rates for Whisper-large with adapters, averaged across 16 languages. Delta indicates the performance disparity between the binary genders.

Results The full set of results is presented in Table 3 of the Appendix, but the summary for Whisper-large – averaging across the 16 languages – is presented in Table 1. We generally see slightly better performance for female speakers than for male speakers. Group-DRO and spectral decoupling perform *on par* with standard ERM fine-tuning, but with the added benefit of lower performance disparity. LoRA also performs comparably, but exhibits higher gender disparity, and Adapter fusion reaches the best performance for both genders, but with higher disparity (Δ) than in Group-DRO and Spectral Decoupling. These trends are observed across all model sizes, but performance is slightly better for larger models, see Figure 2 for performance examples across the 5 Whisper models on English and German VoxPopuli.

4 Discussion

Fairness and (two flavors of) regularization We saw that LoRA exhibits *higher* performance disparities than *vanilla* empirical risk minimization. Since low-rank adaptation is a form of regularization, this seems at odds with previous work suggesting that regularization *mitigates* performance disparities (Sagawa et al., 2020; Petren Bach Hansen et al., 2022). The incongruency is only apparent: It is well-known that sparsity-promoting regularization – like LoRA – hurts robustness (Sutton et al., 2006; Globerson and Roweis, 2006; Søgaard, 2013), while ℓ_2-regularization, ℓ_∞-regularization, noise injection (Bishop, 1995), and drop-out (Wager et al., 2013) often improve robustness. The work cited above (Sagawa et al., 2020; Petren Bach Hansen et al., 2022) both use ℓ_2-regularization.

Rawlsian fairness Whether fairness is best measured by the absolute performance of the worst-off group (Rawlsian fairness) or by the relative performance differences across groups (egalitarian fairness) is a philosophical question (Jørgensen and Søgaard, 2023), but we note that in our case, this has direct consequences for what approach to recommend: Group-DRO or Adapter Fusion? Adapter fusion has superior performance, also for the worst off group, so it is preferable on Rawls' account. The egalitarian would prefer something like Group-DRO, however, since cross-group differences (Δ) are smaller. Note that minimizing cross-group differences is preferable even to *leveling down* (Parfit, 2002) from an egalitarian perspective.

Linguistic fairness Fairness is usually measured across social groups defined by the Cartesian product of a set of protected attributes. However, global technologies can serve some language communities better than others (Lent et al., 2021; Wang et al., 2022; Paz, 2014). This goes for both error rates and disparities (Cabello Piqueras and Søgaard, 2022;

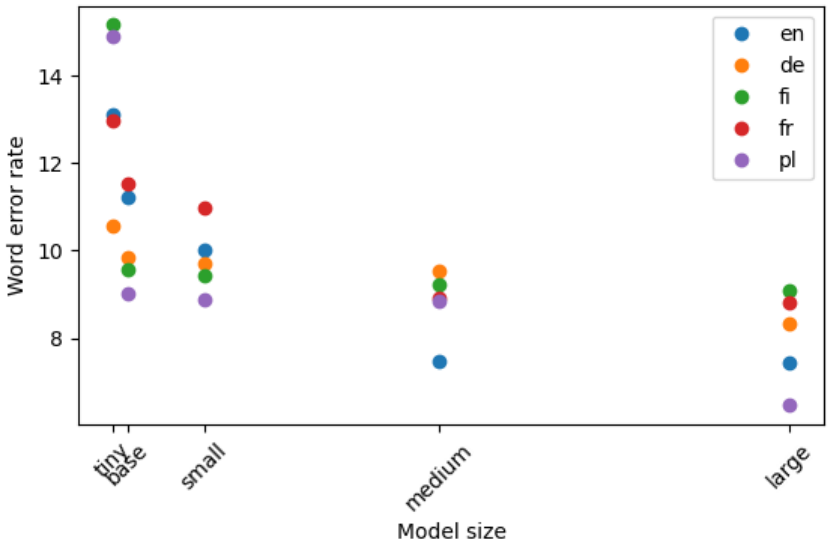

Figure 2: Word error rates with our best performing architecture, adapter fusion, on five languages over model sizes (x-axis).

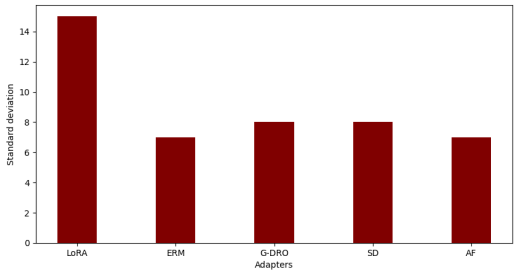

Figure 3: Standard deviations for performance across languages

Ramesh et al., 2023). If Whisper or other multilingual ASR models consistently exhibit much higher error rates or disparities on data in specific languages, their speakers will be disadvantaged. The standard deviations of performances across 16 languages for Whisper-large are given in Figure 3. Fine-tuning cuts more than half of the standard deviation across languages. The fact that adapter fusion leads to the lowest disparity across languages, is promising. Cross-lingual performance gaps such as the ones observed here are still worrying and call for more research; see also Rust and Søgaard (2023) for discussion.

The pros and cons of stacked architectures Adapter fusion adds additional layers of parameters to adjudicate between the three adapter layers trained with different objectives. If we add parameters, we slow down inference time, but on the upside, fine-tuning adapter layers is inexpensive in comparison to fine-tuning encoder or decoder layers of the original Whisper model. Stacking is cheaper than voting in this respect, since we do not have to fine-tune with three different objectives. Finally, the slow inference time can be improved by adapter pruning (Rücklé et al., 2021), without compromising fairness.

5 Conclusions

In this paper, we evaluated the impact on performance and gender disparity of different fine-tuning algorithms in automated speech recognition, including fairness-promoting ones. Our analysis considered model size, language, gender. In general, we saw better performance for female speakers for larger models and significant performance gaps across languages. Fairness-promoting fine-tuning algorithms (Group-DRO and spectral decoupling) hurt performance with some improvements in performance parity. LoRA increases disparities slightly. We find the highest performance with acceptable parity in a novel, fairness-promoting variant of adapter fusion, which had positive ef-

fects on performance, group fairness, as well as parity in performance between languages.

Limitations

While we cover 16 languages, our study is limited to mostly Indo-European, higher-resource languages. English, Slovene, Lithuanian, Italian, French, Polish, Romanian, German, Dutch, Croatian, Slovak, Czech, and Spanish are Indo-European. Hungarian, Finnish, and Estonian are Finno-Ugric languages. It is important to extend studies such as ours to more language families; see, e.g., Abraham et al. (2020). Our dataset only contains binary gender (M/F), and as a result our results are limited to these genders only. While we compare two fairness-promoting fine-tuning strategies, we leave out others, e.g., automatic, worst-case aware curriculum learning (de Lhoneux et al., 2022). Such algorithms could also be combined using adapter fusion. Obviously, it would be relevant to replicate our findings on other multilingual ASR datasets, and it is extremely important to extend studies such as ours to more demographic variables (race, age, language proficiency, impairments, etc.) and to investigate the performance on intersections of these attributes.

References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826, Marseille, France. European Language Resources Association.

Chris M. Bishop. 1995. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116.

Laura Cabello Piqueras and Anders Søgaard. 2022. Are pretrained 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.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–587, Dublin, Ireland. Association for Computational Linguistics.

Amir Globerson and Sam Roweis. 2006. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 353–360, New York, NY, USA. Association for Computing Machinery.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Anna Katrine Jørgensen and Anders Søgaard. 2023. Rawlsian ai fairness loopholes. *AI and Ethics*, 3:1185–1192.

Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. 2020. FACT: A diagnostic for group fairness trade-offs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5264–5274. PMLR.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online. Association for Computational Linguistics.

Derek Parfit. 2002. Equality or priority? In Matthew Clayton and Andrew Williams, editors, *The Ideal of Equality*, pages 81–125. Palgrave Macmillan.

Moria Paz. 2014. The Tower of Babel: Human Rights and the Paradox of Language. *European Journal of International Law*, 25(2):473–496.

Victor Petren Bach Hansen, Atula Tejaswi Neerkaje, Ramit Sawhney, Lucie Flek, and Anders Sogaard. 2022. The impact of differential privacy on group disparity mitigation. In *Proceedings of the Fourth*

Workshop on Privacy in Natural Language Processing, pages 12–12, Seattle, United States. Association for Computational Linguistics.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. 2021. Gradient starvation: A learning proclivity in neural networks. In *Advances 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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phillip Rust and Anders Søgaard. 2023. Differential privacy, linguistic fairness, and training data influence: Impossibility and possibility theorems for multilingual language models. In *International Conference on Machine Learning*.

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. In *International Conference on Learning Representations (ICLR)*.

Anders Søgaard. 2013. Part-of-speech tagging with antagonistic adversaries. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–644, Sofia, Bulgaria. Association for Computational Linguistics.

Charles Sutton, Michael Sindelar, and Andrew McCallum. 2006. Reducing weight undertraining in structured discriminative learning. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 89–95, New York City, USA. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Stefan Wager, Sida Wang, and Percy S Liang. 2013. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Jialu Wang, Yang Liu, and Xin Wang. 2022. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2681–2695, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

			FUSION			ERM	G-DRO	SD	Epochs
Language	learning rate	drop out	lambda_erm	lambda_dro	lambda_spectral	lambda_erm	lambda_dro	lambda_spectral	
English	1.00E-06	0.05	0.8	0.8	0.003	0.5	0.5	0.003	30
German	1.00E-06	0.05	0.8	0.8	0.003	0.5	0.5	0.003	30
French	1.00E-06	0.05	0.8	0.8	0.003	0.5	0.5	0.003	30
Spanish	1.00E-06	0.05	0.8	0.8	0.003	0.5	0.5	0.003	30
Italian	1.00E-06	0.05	0.8	0.8	0.03	0.5	0.5	0.03	30
Romanian	1.00E-05	0.05	0.9	0.9	0.04	0.5	0.5	0.03	30
Hungarian	1.00E-05	0.05	0.9	0.9	0.04	0.5	0.5	0.04	30
Polish	1.00E-04	0.05	0.8	0.8	0.04	0.5	0.5	0.04	30
Czech	1.00E-05	0.05	0.8	0.8	0.04	0.5	0.5	0.03	30
Dutch	1.00E-04	0.05	0.8	0.8	0.04	0.5	0.5	0.04	30
Finnish	1.00E-05	0.05	0.8	0.8	0.03	0.5	0.5	0.04	30
Croatian	1.00E-056	0.05	0.8	0.8	0.03	0.5	0.5	0.04	30
Slovak	1.00E-05	0.05	0.8	0.8	0.03	0.5	0.5	0.04	30
Slovene	1.00E-04	0.05	0.8	0.8	0.03	0.8	0.8	0.03	40
Estonian	1.00E-06	0.05	0.8	0.8	0.05	0.8	0.8	0.05	40
Lithuanian	1.00E-04	0.05	0.8	0.8	0.03	0.8	0.8	0.04	40

Table 2: Hyper-parameters.

Language	LoRA			ERM			G-DRO			SD			Fusion		
English	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	17.85	13.24	15.43	12.79	12.42	12.52	12.47	12.15	12.23	13	12.21	12.66	13.52	12.96	13.12
base	16.39	13.98	15.18	10.45	11.99	11.23	10.43	13.22	11.9	12	13.1	12.5	11.01	11.44	11.23
small	13.58	13.3	13.44	10.41	10.31	10.3	12	11.56	11.8	11.43	11.04	11.33	10.11	10	10
medium	9.31	9.2	9.26	8.06	8.11	8.09	8.1	8.24	8.18	8.66	9.57	9.12	7.37	7.51	7.46
large	10.05	8.49	8.96	7.32	7.45	7.37	7.65	8	7.9	7.88	8.01	7.96	7.16	7.47	7.44
Slovene	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	35.31	40.21	37.8	21.22	26.43	26	25.31	26.11	25.8	26	27.11	27	24.02	28.11	26
base	22.34	36.73	29.6	19.56	23.16	22	21	23.11	22	21.13	24.11	22.66	20	27.11	24
small	21.28	35.3	28.9	20	23.78	21.9	18.56	18.89	18.73	18.46	18.54	18.66	18.75	25	21.9
medium	21.19	33.45	27.3	17.45	18.86	18.17	17.64	18.9	18.28	17.63	18.79	18.22	12.85	13.31	13.16
large	20.15	20.55	20.25	12.17	12.56	12.36	12.56	13.15	12.78	12.48	13.22	12.82	12.12	12.41	12.26
Lithuanian	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	71.42	69.67	70.48	50	50.12	50	51.1	53	52.22	49.91	53.2	51.6	51.12	55.34	53.23
base	59.67	65.71	62.31	36.12	41	39	36	40.19	38	37.03	41.1	39	35.21	41.1	37.78
small	49.34	50.95	50.6	26.76	34.16	30.66	35.09	36.07	35.76	35.09	36.16	35.72	26.43	33.36	30
medium	17.19	19.43	18.4	16.59	17.67	17.11	17.42	17.51	17.47	16.25	17.55	17	16.75	17.55	16.34
large	17.34	18.11	17.6	13.63	14.14	13.89	13.75	14.23	14	13.75	14.24	14	12.13	12.79	12.58
Italian	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	20.42	19.34	20	18	17.15	17.6	17.86	18	17.94	18.93	19.19	19.08	15.24	16.11	15.66
base	16.71	19.12	17.92	17.03	17.54	17.26	16.27	17.25	16.66	18.72	19.04	18.78	14.72	15.11	14.86
small	13.03	14.88	14.54	11.86	13.75	12.95	12.03	13.45	12.67	13.87	14.15	14.01	12.35	13.56	13.25
medium	12.92	14.78	14.45	10.33	11.13	10.8	10.66	11.23	11	10.5	11.02	10.8	10.26	11	10.55
large	8.61	16.21	12.52	8.56	10	9.3	9	10.15	9.6	8.97	10.76	10	7.35	9.96	8.7
French	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	12.54	13.67	14.37	12.32	12.92	12.77	11.8	12.66	12.44	11.57	12.75	12.46	11.46	12.13	12.96
base	11.49	12.27	12.01	11.16	11.54	11.36	11.75	12.08	11.91	11.75	12.08	11.91	10.94	11.86	11.52
small	11.5	12.13	11.81	11	11.42	11.24	10.98	11.54	11.3	11.59	12.03	11.82	10.59	11.34	10.96
medium	9.98	9.72	9.86	8.99	9.43	9.21	9	9.27	9.17	9.12	9.31	9.22	8.75	9.16	8.9
large	9.5	9.63	9.57	8.73	9.1	9	8.99	9.31	9.14	9.01	9.22	9.13	8.56	9.02	8.8
Polish	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	15.8	15.06	15.11	13.45	13.76	13.6	13.72	14.19	13.67	13.7	14.28	13.9	14.91	14.89	14.9
base	11.24	11.46	11.29	9.21	9	9.11	10.14	9.08	9.24	9.48	11.48	10.36	8.91	8.93	9
small	9.8	9.37	9.59	8.55	8.89	8.7	8.56	9	8.9	8.91	9	8.96	8.75	9	8.87
medium	9.78	9.28	9.54	8.51	8.45	8.49	9	8.52	8.84	9	8.75	8.88	8.61	8.5	8.85
large	9.57	9.03	9.31	8.45	8.33	8.4	8.91	8.43	8.66	9.14	8.79	8.87	6.33	6.17	6.46
Romanian	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	48.01	52.71	50.34	28.14	33	30.5	29.19	30.33	30	29	31.12	30	24.67	30.21	27.45
base	29.76	33.17	31.5	16.12	16.31	16.22	16.06	16.11	16.08	16.24	16.35	16.3	15.67	17.03	16.32
small	27.44	27.96	29.23	12.25	12.58	12.41	12.66	13.11	12.89	12.61	13.54	13.07	12.18	13	12.6
medium	9.1	10.71	10	9.81	10	9.9	9.75	10.12	9.94	9.75	10.12	9.94	9	10.23	9.62
large	8.98	9.05	9.22	7.52	8.93	8.55	8.13	9.48	8.81	8.24	9.51	8.86	7.74	8.27	8
German	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	10.72	13.62	11.57	10.13	12.2	11.14	10.21	12.22	11.22	11.12	12.23	11.66	9.65	11.45	10.55
base	10.65	16.63	13.34	9.68	10.44	10.07	9.59	10.63	10.13	10.14	10.44	10.3	9.14	10.55	9.85
small	10.47	12.88	11.7	9.55	10.32	9.94	9.43	10.83	10.21	9.43	10.96	10.2	9.09	10.32	9.71
medium	9.55	10.8	10.17	9.34	10.25	9.8	9.31	10.76	10.06	9.61	10.65	10.15	8.75	9.83	9.54
large	9.26	10.55	9.9	8.73	9.25	9	9.21	9.78	9.51	9.34	9.98	9.68	8.07	8.46	8.32

Table 3: Word error rates (male, female, all) for five fine-tuning algorithms. D-GRO and SD are fairness-promoting. Fusion is our contribution.

Language	LoRA			ERM			G-DRO			SD			Fusion		
Finnish	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	16	17.23	16.62	15	13.55	15.27	14.3	15.34	15.4	15.4	15.64	15.6	15.85	14	15.16
base	10.63	8.89	9.8	10.43	8.55	9.46	10.61	8.75	9.68	10.66	9	9.83	10.23	8.77	9.55
small	10.48	8.83	9.66	10.38	8.32	9.44	10.36	9	9.61	10.4	8.73	9.59	10.16	8.74	9.44
medium	10.22	8.83	9.52	9.47	9.34	9.4	9.43	9.27	9.31	9.47	9.17	9.25	9.44	9.01	9.23
large	10.12	8.56	9.4	9.17	8.66	8.91	9	8.9	8.95	9	8.88	8.95	9.46	8.48	9.08
Dutch	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	21.17	24.35	22.04	20.53	23.24	22.54	19.78	19.32	19.52	19.37	19.17	19.31	16.52	17.04	16.68
base	16.78	19.81	18.3	11.25	12.57	11.91	10.91	11	10.96	10.98	11.71	11.35	11.23	12.61	11.9
small	15.11	16.21	15.66	10.71	11.13	10.93	10.54	11.12	10.8	10.54	11.12	10.8	10.65	11	10.85
medium	11.65	12.74	12.2	9.81	9.99	9.9	9.73	10	9.86	9.81	10.19	10	9.98	10.12	10
large	10.98	11.15	11	9.65	9.78	9.71	8.89	9	8.94	9.19	10	9.6	9.56	10.01	9.79
Croatian	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	17.71	17.88	17.8	16.65	18.07	17.37	15.53	16.24	15.88	15.16	17	16.08	17	17.65	17.33
base	16.32	17.72	17	16.21	17	16.6	15.45	16.14	15.78	15.45	16.14	15.78	16.15	16.55	16.38
small	16.39	17.51	16.76	16.1	17	16.55	14.39	16.18	15.3	15	16.41	15.7	13.45	16.95	14.92
medium	16.34	17	16.67	15.19	16.3	16.38	15.18	15.14	15.18	15.35	16	15.5	13.35	15.46	14.46
large	15.92	16.98	16.21	11.75	12.21	12	12.05	13.24	13	12.05	13.24	13	12.32	13.12	12.8
Hungarian	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	48.46	54.06	50.49	34.72	41.68	36.41	38.06	41.07	38.24	42.06	53.17	42.14	43.41	46.24	44.34
base	27.67	26.76	27.21	20	19.78	19.89	20	21.31	20.65	19	19.81	19.4	21.18	22.57	21.8
small	19.76	18.78	19.3	19.56	19.9	19.74	19.67	20	19.39	19.01	19.56	19.28	18.18	19	18.6
medium	19.34	18.17	21.31	14.31	13.44	13.43	13.55	13.29	13.32	14.12	13.75	13.65	14.51	14.44	14.48
large	12.32	11.46	11.9	10.64	10.55	10.6	10.6	10.47	10.54	10.79	10.66	10.75	10.55	10.21	10.33
Slovak	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	22.21	21.9	22.2	20.18	19.81	19.87	21.18	24.16	20.65	22.35	26.13	25.12	22.01	20.35	21.39
base	26.71	27.46	26.23	18.87	23.41	21.23	23.9	22.34	22.82	22.98	22.31	22.51	18.76	22.49	20.74
small	25.34	25.59	25.41	18.46	19.01	18.74	17.82	20.19	19	17.89	21.16	19.55	17.46	18.1	17.8
medium	16.16	17.32	16.8	13.99	14.41	14.32	15.19	16	15.6	15.39	16	15.8	14.52	16	15.43
large	12.39	14.41	13.79	10.38	10.42	10.41	10.26	9.56	9.8	10.57	12	11.3	8.87	9.41	9.25
Czech	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	26.15	30.62	29.42	14.96	15.58	15.41	15.11	16.11	16.15	16.04	16.65	16.35	24.61	29.85	27.24
base	15.41	24.78	20.62	15.35	16.13	15.75	16	16.1	16.05	16.19	16.47	16.34	15.3	16.03	15.29
small	18.12	19.65	18.51	14.9	15.61	15.26	15	15.43	15.22	15.01	15.83	15.44	14.81	15.79	15.26
medium	15.84	18.81	17.33	15	15.43	15.25	14.99	15.93	15.47	15	16.01	15.51	14.86	15.37	15.23
large	15.71	18.61	17.18	11.69	12.61	12.2	12	13.19	12.6	12.15	13.04	12.6	11.44	12.31	11.89
Spanish	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	10.59	10.58	10.59	10.21	10	10.11	10.24	10.13	10.18	10.24	10.13	10.18	10.19	9.98	10.09
base	10.7	9.91	10.31	9.51	9.07	9.3	9.91	9.31	9.66	9.81	9.32	9.57	9.12	8.91	9.01
small	10.75	9.21	10	9.44	9.21	9.3	9.74	9.52	9.65	9.66	9.02	9.36	10	8.75	8.87
medium	10.72	8.85	9.31	8.54	10.21	9.38	9.49	8.89	9.2	9.5	9	9.27	9.21	8.31	8.66
large	10.68	8.71	9.24	9.41	8.61	9.02	9.24	8.64	8.94	9.34	8.44	9	9.01	8.66	8.49
Estonian	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂	♀	♂	♀+♂
tiny	55.35	54.06	56.09	33.01	28.52	31	30	32.41	31.21	32.14	33.54	32.9	23.26	21.22	22.25
base	55.37	54.13	54.13	19.75	17.43	18.98	33.34	29.24	32.07	32.15	29.91	31.29	13.72	13.52	13.65
small	43.75	41.69	42.46	20.59	18.89	19.7	31.15	28.34	29.8	31.82	29	30.42	13.96	12.21	13.37
medium	20	19.17	19.6	15.47	15.31	15.4	16.16	15.68	16.01	16.16	15.14	15.66	12.66	11.35	12.04
large	18	19.8	19.41	12.94	14.48	13.96	12.03	14.55	13.31	14.56	16	15.3	11.07	12.3	11.89

Table 4: Word error rates (male, female, all) for five fine-tuning algorithms. D-GRO and SD are fairness-promoting. Fusion is our contribution.