Bridging Fairness and Environmental Sustainability in Natural Language Processing

Marius Hessenthaler[1], Emma Strubell[2], Dirk Hovy[3], Anne Lauscher[4]
[1]Data and Web Science Group, University of Mannheim, Germany
[2]Language Technologies Institute, Carnegie Mellon University, U.S.
[3]MilaNLP, Bocconi University, Italy
[4]Data Science Group, University of Hamburg, Germany
marius.hessenthaler@web.de, strubell@cmu.edu,
dirk.hovy@unibocconi.it, anne.lauscher@uni-hamburg.de

Abstract

Fairness and *environmental impact* are important research directions for the sustainable development of artificial intelligence. However, while each topic is an active research area in natural language processing (NLP), there is a surprising lack of research on the interplay between the two fields. This lacuna is highly problematic, since there is increasing evidence that an exclusive focus on fairness can actually hinder environmental sustainability, and vice versa. In this work, we shed light on this crucial intersection in NLP by (1) investigating the efficiency of current fairness approaches through surveying example methods for reducing unfair stereotypical bias from the literature, and (2) evaluating a common technique to reduce energy consumption (and thus environmental impact) of English NLP models, knowledge distillation (KD), for its impact on fairness. In this case study, we evaluate the effect of important KD factors, including layer and dimensionality reduction, with respect to: (a) performance on the distillation task (natural language inference and semantic similarity prediction), and (b) multiple measures and dimensions of stereotypical bias (e.g., gender bias measured via the Word Embedding Association Test). Our results lead us to clarify current assumptions regarding the effect of KD on unfair bias: contrary to other findings, we show that KD can actually *decrease* model fairness.

1 Introduction

Fairness and environmental sustainability are critical to the future of human society, and, thus, also reflected by the United Nations' 17 Sustainable Development Goals (e.g., *Goal 5: Gender Equality*, and *Goal 13: Climate Action*).[1] Accordingly, both topics are currently also active research areas in natural language processing (NLP).

On the one hand, several works have established that language representations are prone to encode and amplify stereotypical social biases (e.g., Bolukbasi et al., 2016), and, consequently, are a source of representational harm (Barocas et al., 2017; Hovy and Spruit, 2016; Shah et al., 2020). To address this issue and provide fairer language technologies, various approaches have developed methods for measuring bias (e.g., Caliskan et al., 2017; Nadeem et al., 2020; Nangia et al., 2020; Nozza et al., 2021, *inter alia*) as well as debiasing methods (e.g., Zhao et al., 2018; Dev and Phillips, 2019, *inter alia*).

On the other hand, recent advances in NLP have been fueled largely by increasingly computationally expensive pre-trained language models (PLMs). Whereas the original BERT base model has 110M parameters (Devlin et al., 2019), the Switch Transformer model, designed as a more efficient alternative to more recent PLMs, has over a trillion parameters (Fedus et al., 2022). While these models consistently obtain superior performance across a variety of NLP benchmarks (Wang et al., 2018, 2019), researchers have pointed out the increasing potential CO_2 emissions of these models. Strubell et al. (2019) estimated that pre-training a BERT base Transformer (Vaswani et al., 2017) using energy with average U.S. carbon intensity has CO_2 emissions comparable to a passenger on a trans-American flight. More recent calculations confirm that the energy consumption of PLMs continues to grow along with their size (Dodge et al., 2022) and that consumption at inference time is non-negligible (Tambe et al., 2021). These findings have fueled the development of more environmentally sustainable NLP. For instance, tuning only a few new lightweight adapter layers instead of the whole architecture (e.g., Houlsby et al., 2019; Pfeiffer et al., 2021), and compressing models (e.g., Gupta and Agrawal, 2022) can reduce the energy consumption during training or inference.

However, while both fairness and sustainability are active research fields in our community,[2] it is

[1] https://sdgs.un.org/goals

[2] See also the proceedings of dedicated workshops,

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7817–7836
December 7-11, 2022

extremely surprising that there is so little work on the intersection of *both* aspects. We argue that **this lack of focus is problematic, as some fairness approaches can jeopardize sustainability and sustainability approaches might hinder fairness**.

For instance, Webster et al. (2020) propose a data-driven debiasing approach, which requires pretraining a fairer model from scratch. Thus, for each and every stereotype, a novel PLM must be trained, reducing environmental sustainability. Lauscher et al. (2021) pointed to potential issues of such an approach and proposed a modular, and therefore more sustainable method. In the other direction, recent work in computer vision has shown that compressed models are less robust, and can even amplify algorithmic bias (Hooker et al., 2020; Liebenwein et al., 2021). Ahia et al. (2021) investigated the relationship between pruning and low-resource machine translation, finding that pruning can actually aid generalization in this scenario by reducing undesirable memorization. However, aside from few works, there has been no systematic research on the interplay between the two fields in NLP.

Contributions. In this work, we acknowledge the potential for race conditions between fairness and environmental sustainability in NLP and call for more research on the interplay between the two fields. To shed light on the problem and to provide a starting point for fair and environmentally sustainable NLP, **(1)** we provide a literature overview and systematize a selection of exemplary fairness approaches according to their sustainability aspects. We show that the surveyed approaches require energy at various training stages and argue that fairness research should consider these aspects. **(2)** Based on work suggesting the potential of model compression to increase fairness (Xu and Hu, 2022), we take a closer look at knowledge distillation (KD; Hinton et al., 2015) as an example method targeting the environmental sustainability of language technology. In this approach, a (smaller) student model is guided by the knowledge of a (bigger) teacher model. We extensively analyze the effect of KD on intrinsic and extrinsic bias measures (e.g., Word Embedding Association Test (e.g., Caliskan et al., 2017), Bias-NLI (Dev et al., 2020)) across two tasks (Natural Language Inference and Semantic Similarity Prediction). We investigate important KD-factors, such as the number of hidden layers of the student and their dimensionality. Contrary to concurrent findings (Xu and Hu, 2022), we show that KD can actually *decrease* fairness. Thus, fairness in such sustainability approaches needs to be carefully monitored. We hope to inspire and inform future research into fair and environmentally sustainable language technology and make all code produced publicly available at: https://github.com/UhhDS/knowledge_distillation_fairness.

2 How Fairness Can Harm Sustainability

To illustrate the tight relationship between environmental sustainability and fairness in current NLP, we conduct an exemplary analysis of current mitigation approaches for unfair bias. Here, our goal is not to conduct an exhaustive survey, but to showcase *when, why, and to what extent* fairness approaches can be environmentally harmful.

2.1 Approach

We query the ACL Anthology[3] for *"debiasing"* and *"bias mitigation"* and examine the first 20 results each. We focus on debiasing of unfair societal stereotypes in monolingual PLMs. Therefore, we exclude approaches on static embeddings, domain generalization,[4] and solely multilingual PLMs. We also consider only papers that propose a novel adaptation or debiasing approach, and exclude papers that survey or benchmark mitigation methods (e.g., Meade et al., 2022). We remove any duplicates.

This approach left us with 8 relevant publications (out of the initial 40 ACL Anthology hits). To diversify the analysis pool, we added one more paper, based on our expert knowledge.

If a paper proposes multiple methods, we focus only on a single method. We apply a coarse-grained distinction between (a) *projection-based*, and (b) *training-based* methods. Projection-based methods follow an analytical approach in a manner similar to the classic hard debiasing (Bolukbasi et al., 2016). In contrast, training-based methods either rely on augmenting training sets (e.g., Zhao et al., 2018) or on a dedicated debiasing loss (e.g., Qian et al., 2019). For the training-based approaches, we additionally classify the stage where the authors demonstrate the debiasing.

e.g., SustaiNLP (https://aclanthology.org/2021.sustainlp-1.0/, and LT-EDI (https://aclanthology.org/2022.ltedi-1.0/))

[3] https://aclanthology.org

[4] As for instance common in the fact verification literature (e.g., Paul Panenghat et al., 2020)

Reference	Method	Type	Increased Environmental Costs?				
			0. Pre-t.	1. Inter.	2. Fine-t.	3. Inf.	Other
Karve et al. (2019)	*Conceptor Debiasing*	Projection	–	–	–	–	🌍
Liang et al. (2020)	*Sent-Debias*	Projection	–	–	–	–	🌍
Kaneko and Bollegala (2021)	*Debias Context. Embs.*	Project. & Train.	–	–	🌍	–	–
Webster et al. (2020)	*Pre-training CDA*	Training	🌍🌍🌍	–	–	–	–
Barikeri et al. (2021)	*Attribute Distance Deb.*	Training	–	🌍🌍	–	–	–
Guo et al. (2022)	*Auto-Debias*	Training	–	🌍🌍	–	–	🌍
Dinan et al. (2020a)	*Biased-controlled Training*	Training	–	🌍🌍	–	–	–
Subramanian et al. (2021)	*Bias-constrained Model*	Training	–	–	🌍	–	–
Lauscher et al. (2021)	*Debiasing Adapters*	Training	–	🌍	(🌍)	🌍	–

Table 1: Overview of examplary debiasing methods w.r.t. their efficiency. We provide information on the type of the approach (*Projection* vs. *Training*), and estimate their environmental impact in 3 classes (🌍– 🌍🌍🌍) in different stages of the NLP-pipeline: *0. Pre-training*, *1. Intermediate Training*, *2. Fine-tuning*, *3. Inference time*, and *Other*.

2.2 Results and Discussion

We show the results of our analysis in Table 1.

Underlying Debiasing Approach. Our small survey yielded examples from a variety of approaches: the *projection-based* approaches are represented by (Karve et al., 2019), (Liang et al., 2020), and (Kaneko and Bollegala, 2021). These require generally only a small amount of energy (🌍) for the analytical computation, which, in some cases, is iteratively applied to improve debiasing performance (Ravfogel et al., 2020). In this case, each iteration will marginally decrease the efficiency. Kaneko and Bollegala (2021) explicitly couple their approach with the model fine-tuning. In contrast, the other 6 works belong to the category of training-based approaches. Here, Webster et al. (2020) and Lauscher et al. (2021) rely on CDA (Zhao et al., 2018) and Dinan et al. (2020a) use control codes to guide the biases. Barikeri et al. (2021) rely on a loss-based bias mitigation for equalizing the distance of opposing identity terms towards sterotypical attributes. Subramanian et al. (2021) use a two-player zero-sum game approach for enforcing fairness constraints and Guo et al. (2022) rely on a prompt-based approach.

Training Stage. For the projection-based approaches, the point in time of their application is not critical to their energy consumption. They can only be applied on a trained model (stages 1–3) and, in general, do not require much energy.

However, for the training-based approaches, the training stage is a vital factor: using them in pretraining (stage 0) corresponds to training a new model from scratch. The energy (and corresponding CO_2 emissions) to perform full PLM pretraining can vary widely. Recent estimates range from 37.3 kWh to train BERT small, to 103.5 MWh to train a 6B parameter Transformer language model (🌍🌍🌍) (Dodge et al., 2022). On the positive side, the model can then be used for a variety of applications without further debiasing, assuming that debiasing transfers (Jin et al., 2021). However, this assumption is under scrutiny (Steed et al., 2022).

Intermediate training requires less energy[5] (🌍🌍) as PLMs have already acquired representation capabilities. However, typically, all parameters are adjusted (e.g., 110M for BERT), and the question of transferability still applies.

Debiasing in the fine-tuning stage seems the most energy efficient (🌍). Still, all parameters must be adjusted and the additional objective and data preparation lead to increased costs. The obvious disadvantage is that for each downstream task and stereotype, debiasing needs to be conducted. Lauscher et al. (2021) propose debiasing adapters. They require less energy in the debiasing procedure (🌍), but add a small overhead at inference time (ca. 1% more parameters). Whether or not they add overhead to the fine-tuning depends on whether developers tune the whole architecture.

Overall, we encourage NLP practitioners to consider the energy efficiency of their debiasing approach in addition to the effectiveness and usability. Energy and emission estimation tools can be used to better estimate the environmental impact of proposed approaches (e.g., Lacoste et al., 2019).

3 How Sustainability Can Harm Fairness

Xu and Hu (2022) hint at the potential of model compression to improve fairness. This finding holds promise for bridging the two fields. Unfortu-

[5] Dodge et al. (2022) report 3.1 kWh to fine-tune BERT small on MNLI, 10x less energy than pre-training

nately, the authors partially use pre-distilled models (for which they cannot control the experimental setup), do not systematically investigate the important dimensions of compression (e.g., hidden size and initialization), and do not address the stochasticity of the training procedure. In contrast, Silva et al. (2021) and Ahn et al. (2022) demonstrate distilled models to be more biased, but either use off-the-shelf models, too, or focus on single bias dimensions and measures only. Gupta et al. (2022) start from the assumption that compression results in unfair models and show it for one setup. We provide the first thorough analysis of compression (using the example of knowledge distillation (KD; Hinton et al., 2015), employing multiple tasks, bias dimensions, and measures) and show that some of these previous assumptions do not hold.

3.1 Knowledge Distillation

The underlying idea of knowledge distillation (KD; Buciluǎ et al., 2006; Hinton et al., 2015) is to transfer knowledge from a (typically big, pre-trained, and highly regularized) *teacher* model to a (typically much smaller and untrained) *student* network. It has been shown that a student network which can learn from the teacher's knowledge is likely to perform better than a small model trained without a teacher's guidance. The knowledge transfer happens through effective supervision from the teacher, e.g., via comparing output probabilities (e.g., Hinton et al., 2015), comparing the intermediate features (e.g., Ji et al., 2021), and initializing the student's layers from the teacher's layers.

3.2 Experimental Setup

Throughout, we use the following setup.

Distillation Tasks, Data Sets, and Measures. We test the effects of KD on two distillation tasks: 1) natural language inference (NLI) using the MNLI data set (Williams et al., 2018), and 2) semantic textual similarity (STS) prediction with the Semantic Textual Similarity-Benchmark (STS-B; Cer et al., 2017) data set. We chose these tasks since they are popular examples of downstream natural language understanding (NLU) tasks. There are also dedicated bias evaluation data sets and measures for the resulting models. For MNLI, we report the accuracy, and for STS the combined correlation score (average of the Pearson's correlation coefficient and Spearman's correlation coefficient).

Fairness Evaluation. Given that some of the existing measures have been shown to be brittle (e.g., Ethayarajh et al., 2019), we ensure the validity of our results by combining *intrinsic* with *extrinsic* measures for assessing stereotypical biases along four dimensions (*gender*, *race*, *age*, and *illness*).

Word Embedding Association Test (WEAT; Caliskan et al., 2017). WEAT is an intrinsic bias test that computes the differential association between two sets of target terms A (e.g., *woman*, *girl*, etc.), and B (e.g., *man*, *boy*, etc.), and two sets of stereotypical attribute terms X (e.g., *art*, *poetry*, etc.), and Y (e.g., *science*, *math*, etc.) based on the mean similarity of their embeddings:

$$w(A,B,X,Y) = \sum_{a\in A} s(a,X,Y) - \sum_{b\in B} s(b,X,Y)\,, \quad (1)$$

with the association s of term $t \in A$ or $t \in B$ as

$$s(t,X,Y) = \frac{1}{|X|}\sum_{x\in X} \cos(\mathbf{t},\mathbf{x}) - \frac{1}{|Y|}\sum_{y\in Y} \cos(\mathbf{t},\mathbf{y})\,. \quad (2)$$

The final score is the effect size, computed as

$$\frac{\mu(\{s(a,X,Y)\}_{a\in A}) - \mu(\{s(b,X,Y)\}_{b\in B})}{\sigma(\{s(t,X,Y)\}_{t\in A\cup B})}\,, \quad (3)$$

where μ is the mean and σ is the standard deviation. To apply the measure, we follow Lauscher et al. (2021), and extract word embeddings from the PLM's encoder, using the procedure proposed by Vulić et al. (2020). We use WEAT tests 3–10[6] which reflect racial (tests 3–5), gender (tests 6–8), illness (test 9), and age bias (test 10).

Sentence Embedding Association Test (SEAT; May et al., 2019). SEAT measures stereotypical bias in sentence encoders following the WEAT principle. However, instead of feeding words into the encoder, SEAT contextualizes the words of the test vocabularies via simple neutral sentence templates, e.g., *"This is <word>."*, *"<word> is here."*, etc. Accordingly, the final score is then based on comparing sentence representations instead of word representations. We use SEAT with the WEAT test vocabularies from tests 3–10, as before. Additionally, we use SEAT's additional Heilman Double Bind (Heilman et al., 2004) Competent and Likable tests which reflect gender bias, and SEAT's Angry Black Woman Stereotype (e.g., Madison, 2009) test, which reflects racial bias.

[6] WEAT tests 1 and 2 consist of bias types which do not consider marginalized social groups (flowers vs. insects, and weapons vs. music instruments)

Bias-STS (Webster et al., 2020). The first extrinsic test is based on the Semantic Textual Similary-Benchmark (STS-B; Cer et al., 2017). The idea is to measure whether a model assigns a higher similarity to a stereotypical sentence pair $s_s = (s_{s1}, s_{s2})$ than to a counter-stereotypical pair $s_c = (s_{c1}, s_{c2})$. Webster et al. (2020) provide templates (e.g., *"A [fill] is walking."*), which they fill with opposing gender identity terms (e.g., *man*, *woman*) and a profession term (e.g., *nurse*) from Rudinger et al. (2018) to obtain 16,980 gender bias test instances consisting of two sentence pairs (e.g., *"A man is walking"* vs. *"A nurse is walking"* and *"A woman is walking"* vs. *"A nurse is walking"*). We train the models on the STS-B training portion and collect the predictions on the created Bias-STS test set. We then follow Lauscher et al. (2021) and report the *average absolute difference* between the similarity scores of male and female sentence pairs.

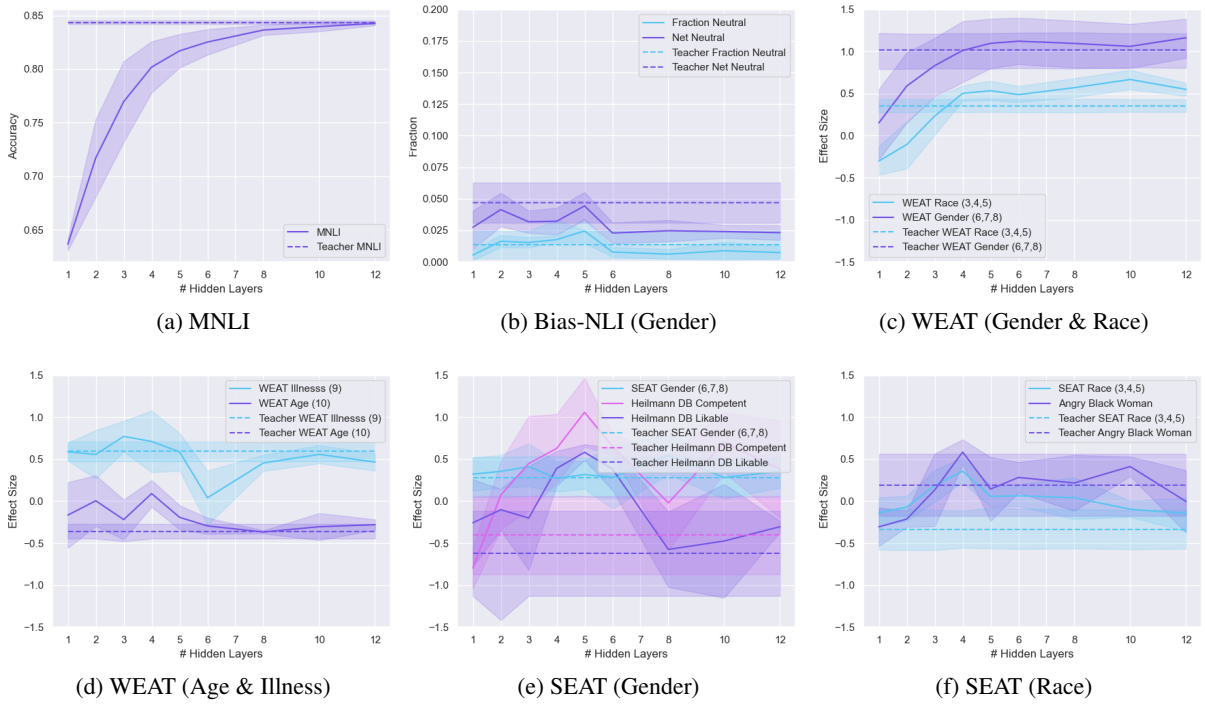
Bias-NLI (Dev et al., 2020). Bias-NLI is another template-based test set, which allows for measuring the tendency of models to produce unfair stereotypical inferences in NLI. We train models on the MNLI training portions, and collect the predictions on the data set. It contains $1,936,512$ instances, which we create using the authors' original code as follows: we start from templates (*"The <subject> <verb> a/an <object>"*) and fill the the verb and object slots with activities (e.g., *"bought a car"*). To obtain a premise we fill the subject slot with an occupation (e.g., *"physician"*), and to obtain the hypothesis, we provide a gendered term as the subject (e.g., *"woman"*). The obtained premise-hypothesis pair (e.g., "*physician bought a car*", "*woman bought a car*") is *neutral*, as we can not make any assumption about the gender of the premise-subject. Accordingly, we can measure the bias in the model with the *fraction neutral* (FN) score — the fraction of examples for which the model predicts the *neutral* class — and as *net neutral* (NN) — the average probability that the model assigns to the *neutral* class across all instances. Thus, in contrast to the other measures, a higher FN or NN value indicates lower bias.

Models and Distillation Procedure. We start from BERT (Devlin et al., 2019) in *base* configuration (12 hidden layers, 12 attention heads per layer, hidden size of 768) available on the Huggingface hub (Wolf et al., 2020).[7] We obtain teacher models from the PLM by optimizing BERT's parameters on the training portions of the respective data sets. We train the models with Adam (Kingma and Ba, 2015) (cross-entropy loss for MNLI, mean-squared error loss for STS-B) for maximum 10 epochs and apply early stopping based on the validation set performance (accuracy for MNLI, combined correlation score for STS-B) with a patience of 2 epochs. We grid search for the optimal batch size $b_t \in \{16, 32\}$ and learning rate $\lambda_t \in \{2 \cdot 10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$. For ensuring validity of our results (Reimers and Gurevych, 2017) we conduct this procedure 3 times starting from different random initializations. As a result, for each of the two tasks, we obtain 3 optimized teacher models. For all distillation procedures, we use the TextBrewer (Yang et al., 2020) framework's *GeneralDistiller*. We optimize the following hyperparameters: batch size $b_d \in \{64, 128\}$ and temperature $t_d \in \{4, 8\}$. We distill for maximum 60 epochs and apply early stopping based on the validation score with a patience of 4 epochs. If we initialize the students' layers, we only apply the task-specific loss on the difference between the teacher's and the student's output. If no layers are initialized, we add a layer matching loss based on Maximum Mean Discrepancy (Huang and Wang, 2017). We use Adam with a learning rate of $1 \cdot 10^{-4}$ (warm up over 10% of the total number of steps and linearly decreasing learning rate schedule).

Dimensions of Analysis. We focus on 3 dimensions: (1) we test the effect of reducing the *number of layers* of the student model and report results on students with 12–1 hidden layers for MNLI and 10–1 hidden layers for STS. All other parameters stay fixed: we set the hidden size to 768 and the number of attention heads per layer to 12 (as in the teacher). We either initialize all layers of the student randomly (for MNLI)[8] or map teacher's layers to student layers for the initialization (for MNLI and STS) according to the scheme provided in the Appendix. (2) The number of layers corresponds to a *vertical* reduction of the model size. Analogously, we study *horizontal* compression reflected by the *hidden size* of the layers. We analyze bias in students with a hidden size $h \in [768, 576, 384, 192, 96]$. Here, we fix the number of hidden layers to 4. We follow Turc et al. (2019) and set the number of self-attention heads

[7] https://huggingface.com

[8] Not mapping the layers, i.e., random initialization, yielded sub par performance for STS



(a) MNLI (b) Bias-NLI (Gender) (c) WEAT (Gender & Race)

(d) WEAT (Age & Illness) (e) SEAT (Gender) (f) SEAT (Race)

Figure 1: Results for our KD analysis (number of student hidden layers) on MNLI without initialization of the layers. We depict (a) the accuracy on MNLI, (b) the fraction neutral and net neutral scores on Bias-NLI, (c) WEAT effect sizes averaged over tests 3–5 (race) and 6–8 (gender), (d) WEAT effect sizes for tests 9 (illness) and 10 (age), (e) SEAT effect sizes averaged over tests 6–8 (gender) and the Heilmann Double Bind tests, and (f) SEAT effect sizes for tests 3–5 (race) and the Angry Black Woman stereotype test. All results are shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and 1–12 layer student models distilled from the teachers.

to $h/64$ and the feed-forward filter-size to $4h$. (3) Finally, we test the effect of the *layer initialization*. To this end, we constrain the student model to have 4 hidden layers, and a hidden size of 768. We then initialize each of the students layers $l_s \in [0, 4]$ (where 0 is the embedding layer) either *individually* or all together with the teacher's layers $l_t \in [0, 12]$ for each experiment with the following mapping ($l_t \rightarrow l_s$): $0 \rightarrow 0, 3 \rightarrow 1$, $6 \rightarrow 2$, $9 \rightarrow 3$, and $12 \rightarrow 4$. For all dimensions, we compare the students' scores with the ones of the teacher model.

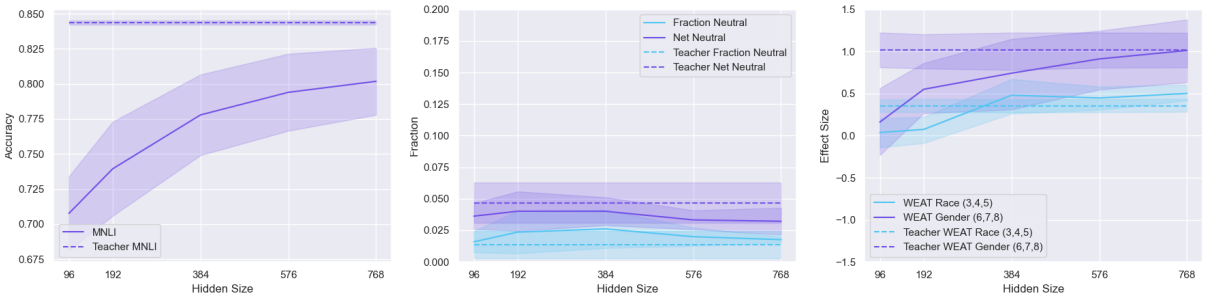
3.3 Results

We discuss the results of our KD analysis.

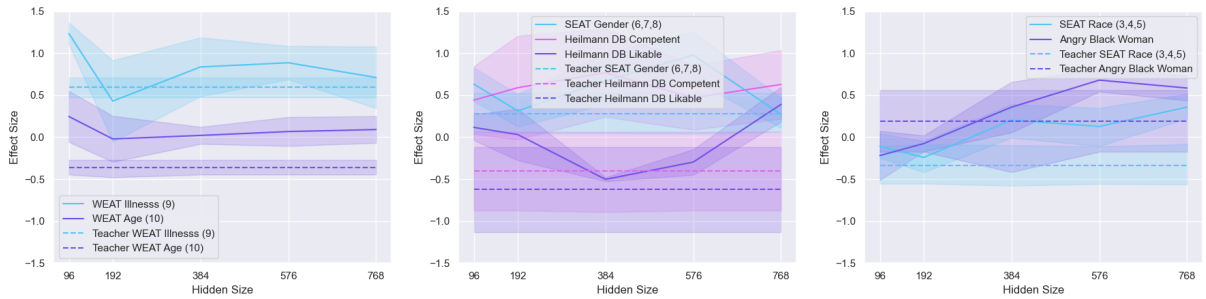
Varying the Number of Hidden Layers. Figures 1a–1f show the MNLI distillation experiments, where we vary the number of student layers (without initializing them). We report the overall performance reflected by MNLI (accuracy) and the bias measured with Bias-NLI, WEAT (Tests 3–10), and SEAT (Tests 3–8, Heilman Double Bind Competent and Likable, and Angry Black Woman Stereotype). We provide the additional SEAT results (Tests 9 and 10) as well as the scores for the other tasks, STS and MNLI with initialization in the Appendix.

The accuracy indicates that we successfully ran the distillation (Figure 1a). Students with 12 hidden layers (no compression) reach roughly the same performance as their teachers. Generally, we observe that the performance variation among students is higher than among teachers, with the highest variation for students with 3 to 5 hidden layers.

Looking at the bias measures (see Figures 1b–1f), we note that the variation of the scores is even higher, especially among the teacher models. This observation suggests lower numerical stability of the bias measures tested. (The test set for Bias-NLI contains ∼2 Million instances, so this aspect cannot be attributed to lower test set sizes). Unsurprisingly, the bias results of the students are generally in roughly the same areas than the ones of their teachers. *This shows that students inherit their teachers biases in the distillation process*. Grouping the test results by measure (e.g., WEAT, etc.) and dimension (e.g., race) results in roughly the same patterns of biases measurable. E.g., in Figure 1f, the results of the aggregated tests 3, 4, and 5 follow the same pattern as the Angry Black Woman

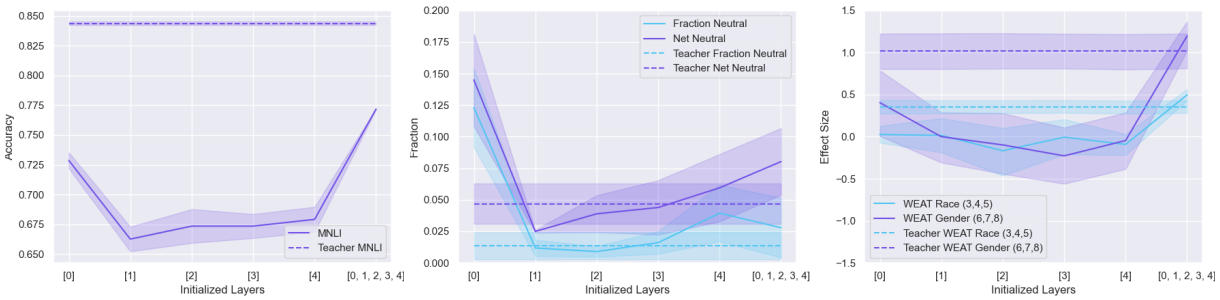


(a) MNLI (b) Bias-NLI (Gender) (c) WEAT (Gender & Racism)



(d) WEAT (Age & Illness) (e) SEAT (Gender) (f) SEAT (Race)

Figure 2: Results for our KD analysis (varying hidden size) on MNLI (without initialization of student layers, 4 hidden layers). We depict (a) the accuracy on MNLI, (b) the fraction neutral and net neutral scores on Bias-NLI, (c) WEAT effect sizes averaged over tests 3,4,5 (race) and 6,7,8 (gender), (d) WEAT effect sizes for tests 9 (illness) and 10 (age), (e) SEAT effect sizes averaged over tests 6–8 (gender) and the Heilmann Double Bind tests, and (f) SEAT effect sizes for tests 3–5 (race) and the Angry Black Woman stereotype test. All results shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and 96–768 hidden size student models.



(a) MNLI (b) Bias-NLI (Gender) (c) WEAT (Gender & Racism)

Figure 3: Results MNLI-KD when varying initialization of the student layers. We depict (a) the accuracy on MNLI, (b) the FN and NN scores on Bias-NLI, and (c) WEAT effect sizes averaged over tests 3,4,5 (race) and 6,7,8 (gender). All results are averages with 90% confidence interval for the 3 teacher models (dashed lines) and students distilled from the teachers where either a single layer was initialized ([0], [1], [2], [3], or [4]) or all layers ([0, 1, 2, 3, 4]).

Stereotype test. We hypothesize that this is due to the partially overlapping term sets. However, across measures and dimensions we find roughly the same bias behavior: students with 12 to 6 hidden layers often exhibit a higher bias than their teachers (for NLI, this corresponds to a lower FN)! The exception to this rule is WEAT test 9, illness. For most tests, the highest bias arises with 4 hidden layers. Students with lower number of layers are mostly less biased across all tests. However, from this point on, the accuracy on MNLI also drops more strongly. These findings are in stark contrast to the results of Xu and Hu (2022).

Varying the Hidden Size. We show the results of KD when varying the hidden size of the student models (number of hidden layers is fixed to 4) in Figures 2a–2f. As before, we provide additional

results in the Appendix. Generally, we note that the performance curve (Figure 2a) is again in-line with our expectations. As in the previous experiment we note high variations of the scores and students biases mostly seem to be located in roughly the same ball park as their teachers' scores. However, we again note that the concrete behavior of the bias curves depends on bias measure and dimension. Interestingly, the curves when varying the hidden size look (with some exceptions) similar to the ones when varying the number of hidden layers. We thus hypothesize, that *both vertical and horizontal compression have a similar affect on fairness*.

Varying the Initialization. As a last aspect of our analysis, we look at the effect of initializing various layers of the student with the weights of the teacher. We depict some of the scores in Figures 3a–3c. Interestingly, changing the initialization has a large effect both on the MNLI accuracy, as well as on the bias measures. These findings highlight again that *monitoring fairness after KD is crucial*.

Overall, our findings show that the devil is in the detail. While generally, **the amount of bias in the distilled models is inherited from the teacher's biases** and **the biases measurable seem to roughly group by social bias dimension and measure**, biases still need to be carefully tested. Most importantly, while Xu and Hu (2022) point at the potential of KD for increasing fairness, we cannot confirm this observation. In contrast, **across most bias measures tested, the student models start from a higher amount of bias than the teacher**. A possible explanation for this behavior is that *weak learners*, i.e., models with limited capacity, generally show a stronger tendency to exploit biases in the data set during the learning process than models with higher capacity (Sanh et al., 2020).

4 Related Work

Fairness in NLP. There exists a plethora of works on increasing the fairness of NLP models, most prominently focused on the issue of unfair stereotypes in the models (e.g., Caliskan et al., 2017; Zhao et al., 2017; Dev et al., 2020; Nadeem et al., 2020, *inter alia*). We only provide an overview and refer the reader to more comprehensive surveys on the topic (e.g., Sun et al., 2019; Blodgett et al., 2020; Shah et al., 2020). Bolukbasi et al. (2016) were the first to point to the issue of stereotypes encoded in static word embeddings, which led to a series of works focused on measuring and mitigating these biases (e.g., Dev and Phillips, 2019; Lauscher et al., 2020a), as well as assessing the reliability of the tests (Gonen and Goldberg, 2019; Ethayarajh et al., 2019; Antoniak and Mimno, 2021; Delobelle et al., 2021; Blodgett et al., 2021). For instance, Caliskan et al. (2017) proposed the well-known WEAT. Recent works focus on measuring and mitigating bias in contextualized language representations (Kurita et al., 2019; Bordia and Bowman, 2019; Qian et al., 2019; Webster et al., 2020; Nangia et al., 2020; Sap et al., 2020) and in downstream scenarios, e.g., for dialog (e.g., Sheng et al., 2019; Dinan et al., 2020a; Barikeri et al., 2021), co-reference resolution (Zhao et al., 2018), and NLI (Rudinger et al., 2017; Dev et al., 2020). Similarly, researchers have explored multilingual scenarios (e.g., Lauscher and Glavaš, 2019; Lauscher et al., 2020c; Ahn and Oh, 2021), more fine-grained biases (Dinan et al., 2020b), and more biases, beyond the prominent sexism and racism dimensions (e.g., Zhao et al., 2018; Rudinger et al., 2018), like speciesist bias (Takeshita et al., 2022).

Sustainability in NLP. Strubell et al. (2019) have called for more awareness of NLP's environmental impact. Reducing the energy consumption can be achieved through efficient pre-training (Di Liello et al., 2021), smaller models and employing less pre-training data considering the specific needs of the task at hand (e.g., Pérez-Mayos et al., 2021; Zhang et al., 2021). If a PLM is already in-place, one can rely on sample-efficient methods (e.g., Lauscher et al., 2020b), or refrain from fully fine-tuning the model (e.g, Houlsby et al., 2019; Pfeiffer et al., 2021). Similarly, one can compress the models via distillation (e.g., Hinton et al., 2015; Sanh et al., 2019; He et al., 2021), pruning (e.g., Fan et al., 2019; Li et al., 2020; Wang et al., 2020), and quantization (e.g., Zhang et al., 2020), to increase energy-efficiency of later training stages or at inference time. A survey is provided by Gupta and Agrawal (2022). In the area of distillation, researchers have explored distillation in different setups, e.g., for a specific task (e.g., See et al., 2016), on a meta-level (e.g., He et al., 2021), or for a specific resource scenario (e.g., Wasserblat et al., 2020). Other efforts focused on accurate energy and emission measurement and provide tools for monitoring energy consumption (e.g., Lacoste et al., 2019; Cao et al., 2020). While most research in the area of NLP focuses on reducing operational costs, i.e., carbon emissions due to the energy re-

quired to develop and run models, downstream impacts of model deployment stand to have a much larger impact on the environment (Kaack et al., 2022). See Rolnick et al. (2022) for a detailed presentation of how machine learning can help to counter climate change more broadly, including a disucussion of NLP applications.

Bridging Fairness and Sustainability. To the best of our knowledge, there are currently only few works that are located at the intersection of the two fields in NLP: Lauscher et al. (2021) proposed to use adapters for decreasing energy consumption during training-based debiasing and increasing the reusability of this knowledge, which has been proven effective by Holtermann et al. (2022). Recently, the unpublished work of Xu and Hu (2022) asks whether compression can improve fairness. In contrast, Silva et al. (2021) find that off-the-shelf distilled models, such as DistilBERT, exhibit higher biases, but do not provide a systematic evaluation of the effect of KD dimensions. Concurrent to our work, Ahn et al. (2022) demonstrate similar trends, but focus on gender bias (quantified through a single measure) and the number of hidden layers in the student, only. Starting from the assumption that compression can lead to biased models, Gupta et al. (2022) propose a fairness-increasing KD loss and demonstrate their baselines to be more biased. In a similar vein, Xu et al. (2021) discuss the robustness of BERT compression. In computer vision, researchers have shown that compression exacerbates algorithmic bias (e.g., Hooker et al., 2020). E.g., Liebenwein et al. (2021) demonstrate pruned models to be brittle to out-of-distribution points. Ahia et al. (2021) present the most relevant work in this space, exploring the *low-resource double-bind*: individuals with the least access to computational resources are also likely to have scarce data resources. They find that model pruning can lead to better performance on low-resource languages by reducing undesirable memorization of rare examples. This study represents a valuable step towards better understanding the intersection of fairness and sustainability. In this work, we argue that more research is needed to understand the complex relationships between the two fields.

5 Conclusion

Fairness and environmental sustainability are equally important goals in NLP. However, the vast majority of research in our community focuses exclusively on one of these aspects. We argue that bridging fairness and environmental sustainability is thus still an unresolved issue. To start bringing these fields together in a more holistic research on ethical issues in NLP, we conducted a two-step analysis: first, we provided an overview on the efficiency of exemplary fairness approaches. Second, we ran an empirical analysis of the fairness of KD, as a popular example of methods to enhance sustainability. We find that use of KD can actually decrease fairness, motivating our plea for research into joint approaches. We hope that our work inspires such research on the interplay between the two fields for fair and sustainable NLP.

Acknowledgements

This work is in part funded by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). The work of Anne Lauscher is funded under the Excellence Strategy of the Federal Government and the Länder. At the time of writing, DH and AL were members of the Data and Marketing Insights unit of the Bocconi Institute for Data Science and Analysis. We thank the anonymous reviewers for their insightful comments.

Limitations

Our work deals with the general relationship between environmental sustainability and fairness. As a showcase, we explore the effect of KD on stereotypical bias measures. This does not imply that KD is the only or the most egregious method, and more research into other approaches is needed. In this context, we resorted to established bias measures, which treat gender as a binary variable. This is due to the limitations of the established data sets, some of which allow for measuring "classic" sexism and do not reflect the large spectrum of possible identities (Lauscher et al., 2022). We further acknowledge that, in this research, we only worked with examples to highlight the importance of considering both sustainability goals. We acknowledge that to truly understand the relationship between fairness and environmental sustainability, it requires more in-depth studies. We thus encourage future research to explore the interplay between the fields more, for other societal biases, including but not limited to queerphobia and non-binary exclusion (Dev et al., 2021), and for other sustainability and fairness approaches and aspects.

References

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. 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 1889–1904, Online. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational 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 1941–1955, Online. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA. Association for Computing Machinery.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Qingqing Cao, Aruna Balasubramanian, and Niranjan Balasubramanian. 2020. Towards accurate and reliable energy measurement of NLP models. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 141–148, Online. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447*.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luca Di Liello, Matteo Gabburo, and Alessandro Moschitti. 2021. Efficient pre-training objectives for transformers. *arXiv preprint arXiv:2104.09694*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the carbon intensity of ai in cloud instances. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Manish Gupta and Puneet Agrawal. 2022. Compression of deep learning models for text: A survey. *ACM Trans. Knowl. Discov. Data*, 16(4).

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

Haoyu He, Xingjian Shi, Jonas Mueller, Sheng Zha, Mu Li, and George Karypis. 2021. Distiller: A systematic study of model distillation methods in natural language processing. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 119–133, Virtual. Association for Computational Linguistics.

Madeline E Heilman, Aaron S Wallen, Daniella Fuchs, and Melinda M Tamkins. 2004. Penalties for success: reactions to women who succeed at male gender-typed tasks. *Journal of applied psychology*, 89(3):416.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv e-prints*, pages arXiv–1503.

Carolin Holtermann, Anne Lauscher, and Simone Ponzetto. 2022. Fair and argumentative language modeling for computational argumentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7841–7861, Dublin, Ireland. Association for Computational Linguistics.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. In *Fifth Workshop on Human Interpretability in Machine Learning (WHI)*.

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 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Zehao Huang and Naiyan Wang. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Mingi Ji, Byeongho Heo, and Sungrae Park. 2021. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7945–7952.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020a. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8131–8138.

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.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020b. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020c. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Bingbing Li, Zhenglun Kong, Tianyun Zhang, Ji Li, Zhengang Li, Hang Liu, and Caiwen Ding. 2020. Efficient transformer-based large scale language representations using hardware-friendly block structured pruning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3187–3199, Online. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence

representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Lucas Liebenwein, Cenk Baykal, Brandon Carter, David Gifford, and Daniela Rus. 2021. Lost in pruning: The effects of pruning neural networks beyond test accuracy. In *Proceedings of Machine Learning and Systems*, volume 3, pages 93–138.

D. Soyini Madison. 2009. Crazy patriotism and angry (post)black women. *Communication and Critical/Cultural Studies*, 6(3):321–326.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Mithun Paul Panenghat, Sandeep Suntwal, Faiz Rafique, Rebecca Sharp, and Mihai Surdeanu. 2020. Towards the necessity for debiasing natural language inference datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6883–6888, Marseille, France. European Language Resources Association.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2022. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2).

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version

of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2020. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. 2016. Compression of neural machine translation models via pruning. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 291–301, Berlin, Germany. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Masashi Takeshita, Rafal Rzepka, and Kenji Araki. 2022. Speciesist language and nonhuman animal bias in english masked language models. *arXiv preprint arXiv:2203.05140*.

Thierry Tambe, Coleman Hooper, Lillian Pentecost, Tianyu Jia, En-Yu Yang, Marco Donato, Victor Sanh, Paul Whatmough, Alexander M. Rush, David Brooks, and Gu-Yeon Wei. 2021. Edgebert: Sentence-level energy optimizations for latency-aware multi-task nlp inference. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '21, page 830–844, New York, NY, USA. Association for Computing Machinery.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the*

2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2020. Structured pruning of large language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6151–6162, Online. Association for Computational Linguistics.

Moshe Wasserblat, Oren Pereg, and Peter Izsak. 2020. Exploring the boundaries of low-resource BERT distillation. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 35–40, Online. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10653–10659, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guangxuan Xu and Qingyuan Hu. 2022. Can model compression improve nlp fairness. *arXiv preprint arXiv:2201.08542*.

Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu. 2020. TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 9–16, Online. Association for Computational Linguistics.

Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. 2020. TernaryBERT: Distillation-aware ultra-low bit BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 509–521, Online. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? 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 1112–1125, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Links to Data, Models, and Code Bases

We provide the links to datasets and code bases used in this work in Table 2. In all distillation experiments, we start from BERT in base configuration: https://huggingface.co/bert-base-uncased. Our code is provided in the GitHub repository linked in the main body of the manuscript.

A.1 Details on the Initialization

When varying the hidden layers of the student and fully initializing the layers, the layer initialization is dependent on the student's size. Following common practice, the mapping is spread across the layers of the teacher. We provide the mapping here ($l_t \rightarrow l_s$):

- 10 student layers: $0 \rightarrow 0$, $2 \rightarrow 1$, $3 \rightarrow 2$, $4 \rightarrow 3$, $5 \rightarrow 4$, $6 \rightarrow 5$, $7 \rightarrow 6$, $8 \rightarrow 7$, $9 \rightarrow 8$, $10 \rightarrow 9$ $12 \rightarrow 10$
- 8 student layers: $0 \rightarrow 0$, $2 \rightarrow 1$, $4 \rightarrow 2$, $5 \rightarrow 3$, $6 \rightarrow 4$, $7 \rightarrow 5$, $8 \rightarrow 6$, $10 \rightarrow 7$, $12 \rightarrow 8$
- 6 student layers: $0 \rightarrow 0$, $2 \rightarrow 1$, $4 \rightarrow 2$, $6 \rightarrow 3$, $8 \rightarrow 4$, $10 \rightarrow 5$, $12 \rightarrow 6$
- 5 student layers: $0 \rightarrow 0$, $3 \rightarrow 1$, $5 \rightarrow 2$, $7 \rightarrow 3$, $9 \rightarrow 4$, $12 \rightarrow 5$
- 4 student layers: $0 \rightarrow 0$, $3 \rightarrow 1$, $6 \rightarrow 2$, $9 \rightarrow 3$, $12 \rightarrow 4$
- 3 student layers: $0 \rightarrow 0$, $4 \rightarrow 1$, $8 \rightarrow 2$, $12 \rightarrow 3$
- 2 student layers: $0 \rightarrow 0$, $6 \rightarrow 1$, $12 \rightarrow 2$
- 1 student layer: $0 \rightarrow 0$, $6 \rightarrow 1$

B Additional Results

We provide the additional results for our distillation experiments.

B.1 Varying the Number of Student Layers

The additional results for the MNLI distillation, i.e., the SEAT scores for tests 9 and 10 are shown in Figure 4. The MNLI results when initializing all layers are depicted in Figure 6.

B.2 Varying the Hidden Size

We provide the additional results for MNLI when varying the student's hidden size (without initialization of student layers, 4 hidden layers) in Figure 5.

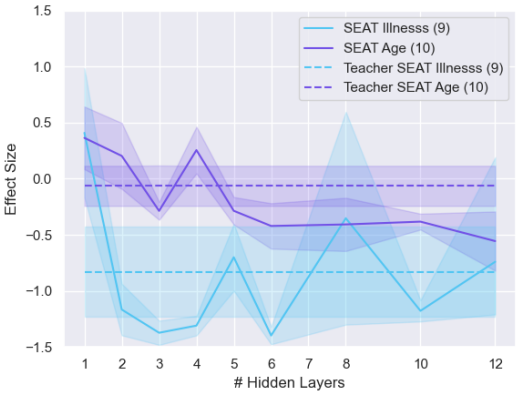


Figure 4: Additional results for our KD analysis (number of hidden layers w/o initialization). We show the MNLI distillation results for SEAT tests 9 and 10.

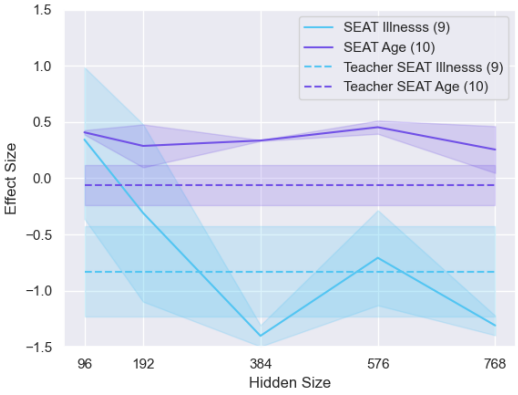


Figure 5: Additional results for our KD analysis (varying hidden size) on MNLI (without initialization of student layers, 4 hidden layers). We show the MNLI distillation results for SEAT tests 9 and 10.

B.3 Varying the Initialization

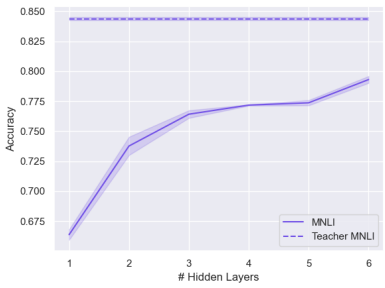
We provide the additional results when varying the initialization for MNLI in Figure 7.

B.4 Semantic Textual Similarity

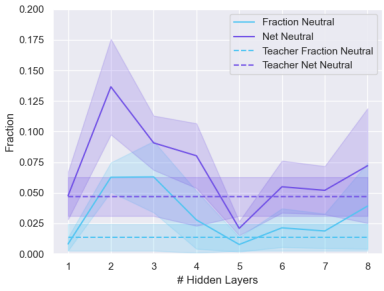
Finally, we also provide the results of our distillation on STS in Figure 8.

Purpose	Name	URL
Natural Language Inference Data	MNLI	https://huggingface.co/datasets/glue
Semantic Similarity Prediction Data	STS-B	https://huggingface.co/datasets/glue
Intrinsic Bias Test Terms	WEAT	https://www.science.org/doi/10.1126/science.aal4230
Intrinsic Bias Test Sentences	SEAT	https://github.com/W4ngatang/sent-bias
Extrinsic Bias Code for Data	Bias-NLI	https://github.com/sunipa/On-Measuring-and-Mitigating-Biased-Inferences-of-Word-Embeddings
Extrinsic Bias Templates	Bias-STS	https://arxiv.org/pdf/2010.06032.pdf
General Code Base	Transformers	https://transformer.huggingface.co
Code Base for Distillation	TextBrewer	https://github.com/airaria/TextBrewer

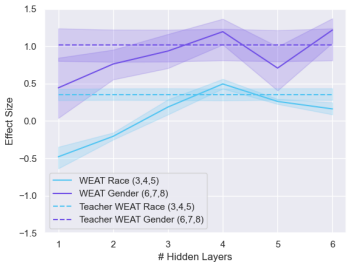
Table 2: Links to the datasets and code bases used in our work.



(a) MNLI

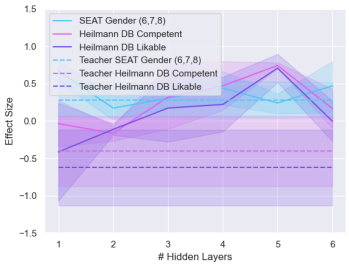


(b) Bias-NLI

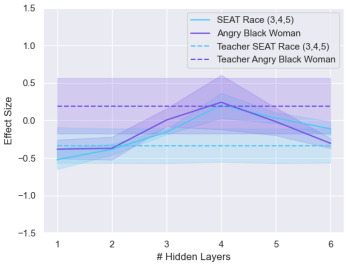


(c) WEAT (Gender & Racism)

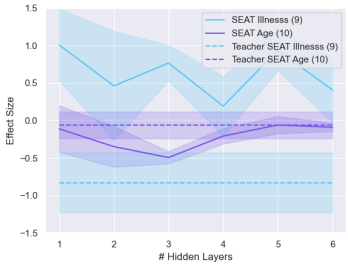
(d) WEAT (Age & Illness)



(e) SEAT (Gender)

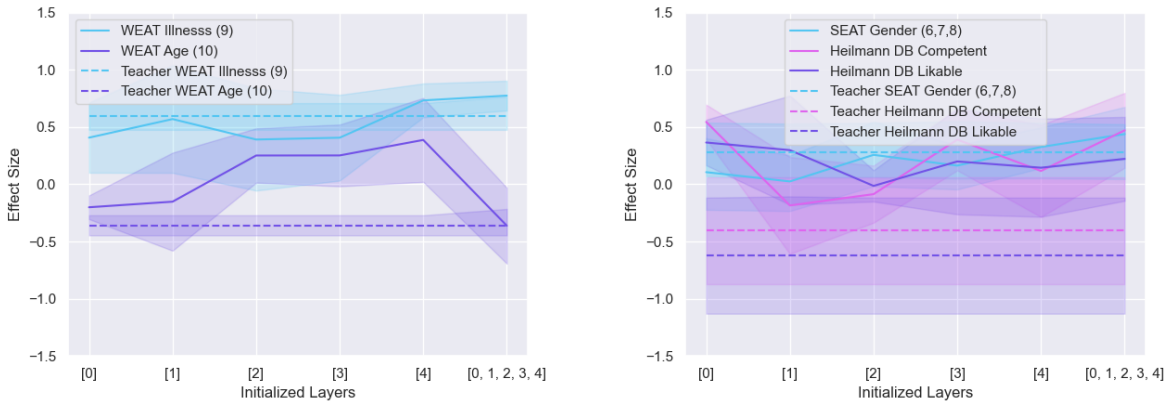


(f) SEAT (Race)



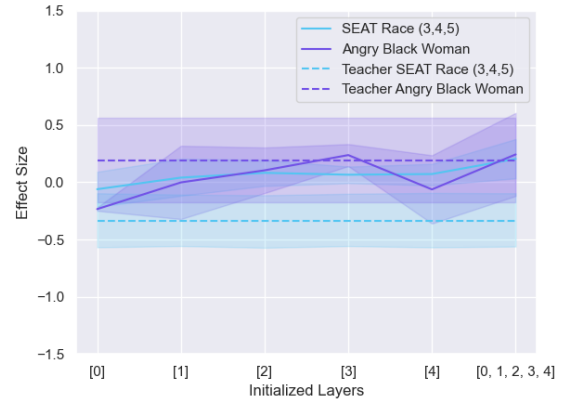
(g) SEAT (Age & Illness)

Figure 6: Results for our KD analysis (number of hidden layers) on MNLI distillation with initialization of all layers. We depict (a) the accuracy on MNLI, (b) the fraction neutral and net neutral scores on Bias-NLI, (c) WEAT effect sizes averaged over tests 3,4,5 (race) and 6,7,8 (gender), (d) WEAT 9 and 10 (age and illness), (e) SEAT scores for tests 6,7,8 and the Heilmann Double Bind tests (gender), (f) SEAT scores for 3,4,5 and Angry Black Woman Stereotype, and (g) SEAT 9 and 10 (age and illness). All results are shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and 1–12 layer student models distilled from the teachers.

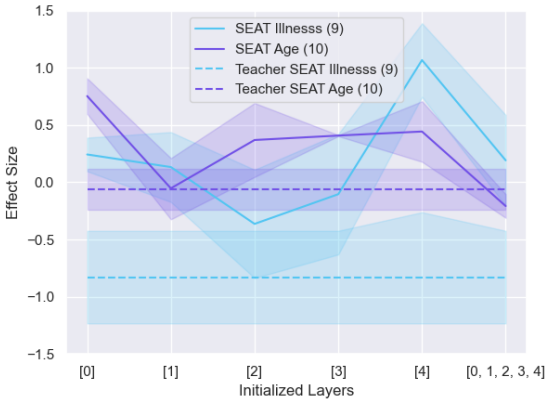


(a) WEAT (Age & Illness)

(b) SEAT (Gender)

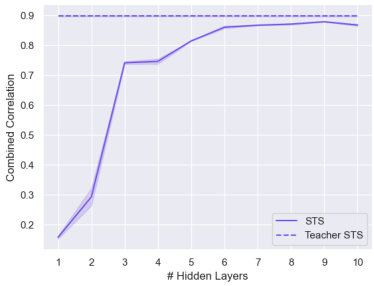


(c) SEAT (Race)

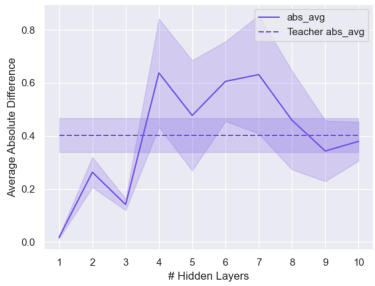


(d) SEAT (Age & Illness)

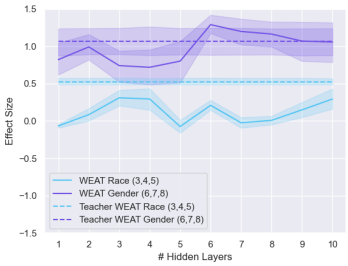
Figure 7: Results for our KD analysis (varying initialization of the student layers) on MNLI. We depict (a) WEAT Age & Illness, (b) SEAT Gender, (c) SEAT Race, and (d) SEAT Age & Illness. All results are shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and student models distilled from the teachers where either a single layer was initialized ([1], [2], [3], or [4]) or all layers ([1, 2, 3, 4]).



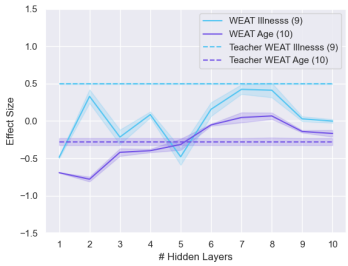
(a) STS-B



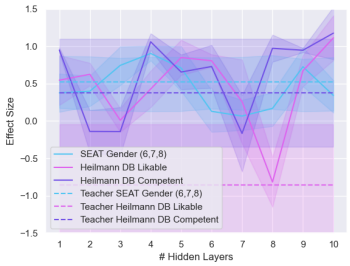
(b) Bias-STS (Gender)



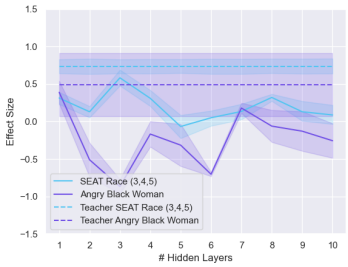
(c) WEAT (Gender & Racism)



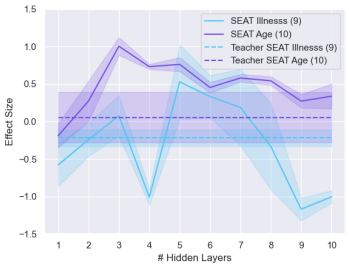
(d) WEAT (Age & Illness)



(e) SEAT (Gender)



(f) SEAT (Race)



(g) SEAT (Age & Illness)

Figure 8: Results for our KD analysis (number of hidden layers) on STS-B distillation with initialization of all layers. We depict (a) the correlation on STS (measured as average of the Pearson and Spearman correlation coefficients), (b) the average absolute difference on Bias-STS, (c) WEAT effect sizes averaged over tests 3,4,5 (race) and 6,7,8 (gender), (d) WEAT 9 and 10 (age and illness), (e) SEAT scores for tests 6,7,8 and the Heilmann Double Bind tests (gender), (f) SEAT scores for 3,4,5 and Angry Black Woman Stereotype, and (g) SEAT 9 and 10 (age and illness) All results are shown as average with 90% confidence interval for the 3 teacher models (dashed lines) and 1–12 layer student models distilled from the teachers.