Do Students Debias Like Teachers?
On the Distillability of Bias Mitigation Methods

Anonymous ACL submission

Abstract

Knowledge distillation (KD) is an effective method for model compression and transferring knowledge between models. However, its effect on model's robustness against spurious correlations that degrade performance on out-of-distribution data remains underexplored. This study investigates the effect of knowledge distillation on the transferability of "debiasing" capabilities from teacher models to student models on natural language inference (NLI) and image classification tasks. Through extensive experiments, we illustrate several key findings: (i) overall the debiasing capability of a model is undermined post-KD; (ii) training a debiased model does not benefit from injecting teacher knowledge; (iii) although the overall robustness of a model may remain stable post-distillation, significant variations can occur across different types of biases; and (iv) we pin-point the internal attention pattern and circuit that causes the distinct behavior post-KD. Given the above findings, we propose three effective solutions to improve the distillability of debiasing methods: developing high quality data for augmentation, implementing iterative knowledge distillation, and initializing student models with weights obtained from teacher models. To the best of our knowledge, this is the first study on the effect of KD on debiasing and its interenal mechanism at scale. Our findings provide understandings on how KD works and how to design better debiasing methods.

1 Introduction

Machine learning models are susceptible to spurious correlations in datasets, also known as "shortcuts" or "dataset biases" (Liu et al., 2015; McCoy et al., 2019). Models that rely on shortcuts can achieve high performance on in-domain or over-represented data, but degrade significantly on out-of-distribution or under-represented data (Li et al., 2023; Chew et al., 2024; Li et al., 2025).

Despite recent advancements in bias mitigation (Guo et al., 2023; Noohdani et al., 2024; Cheng and Amiri, 2024a; Bombari and Mondelli, 2025) and knowledge distillation (Stanton et al., 2021; Sultan, 2023; Sun et al., 2024; He et al., 2025), the effect of knowledge distillation on debiasing at scale is largely unexplored. The internal mechanisms causing that effect remains unclear. This work studies the following research questions (RQs):

- **RQ1**: *To what extent can knowledge distillation transfer debiasing capabilities between models?*
- **RQ2**: *Can knowledge distillation train less biased models compared to standard training?*
- **RQ3**: *What internal mechanisms cause the debiasing behavior change after distillation?*

Answering these questions will help us understand the efficacy of knowledge distillation in handling dataset biases, its underlying mechanisms, and its role in developing new training methods for bias mitigation at scale.

We answer these questions by designing and conducting an empirical analysis on natural language understanding and image classification tasks. Our analyses show that: (i) the effect of knowledge distillation on debiasing performance depends on the underlying debiasing method, the relative scale of the models involved, and the size of the training set; (ii) knowledge distillation effectively transfers debiasing capabilities when teacher and student are similar in scale (number of parameters); (iii) knowledge distillation may amplify the student model's reliance on spurious features, and this effect does not diminish as the teacher model scales up; and (iv) although the overall robustness of a model may remain stable post-distillation, significant variations can occur across different types of biases; and (v) consistent

transfer patterns sometimes emerge, such as performance gap between teacher and student on out-of-distribution (OOD) data, suggesting the possibility of predictable changes in robustness after distillation. Given the above findings, we propose three effective solutions to improve the distillability of debiasing methods: developing high quality data for augmentation, implementing iterative knowledge distillation, and initializing student models with weights obtained from teacher models.

We summarize our contributions as follows:

- we present the first study (to the best of our knowledge) of the effect of knowledge distillation on dataset bias at scale across both language and vision tasks;
- we investigate the internal mechanisms that causes debiasing ability change before and after distillation, namely the divergence of attention and change of circuit;
- we propose three strategies to improve the distillability of debiasing methods and provide insights for future development of bias mitigation techniques.

2 Knowledge Distillation and Debiasing

Problem Formulation We investigate the effect of knowledge distillation (KD) on debiasing methods. Denote $\mathcal{A}$ as a debiasing training algorithm to train robust models agains spurious correlations, and m as a model. We are interested in the *distillability of debiasing methods* the extent of debiasing that retains after knowledge distillation, i.e.

$$D = \mathcal{A}(m) - \text{KD}(\mathcal{A}(m)). \quad (1)$$

We are also interested in the *contribution of KD* – the performance improvement gained by training a debiasing method with KD over training without KD, i.e.

$$C = \mathcal{A}(m) - \mathcal{A}(m; \text{KD}). \quad (2)$$

Notation and Training Setup Let f and g denote models trained without knowledge distillation and with knowledge distillation respectively. In this paper, we use subscript $\mathcal{T}$ and $\mathcal{S}$ to denote teacher and student scales respectively. As illustrated in Figure 1, we train the following models for each debiasing method: **(i)** we train from scratch for both teacher and student scales to obtain $f_\mathcal{T}$ and $f_\mathcal{S}$, see Figure 1(a). **(ii)** Then for every scale $\mathcal{T} > \mathcal{S}$, we distill the knowledge from $f_\mathcal{T}$ to $g_{\mathcal{T}->\mathcal{S}}$, see Figure 1(b). Given a debiasing method M and the three models obtained above ($f_\mathcal{T}$, $f_\mathcal{S}$, and $g_{\mathcal{T}->\mathcal{S}}$), we conduct the following comparisons:

- **C1: Teacher ($f_\mathcal{T}$) vs. Student ($g_{\mathcal{T}->\mathcal{S}}$).** This comparison reveals if knowledge distillation can distill debiasing capability between models and if it affects model's robustness to spurious correlations, which answers *RQ1* (§4.1).
- **C2: Non-KD vs. KD**, realized by comparing $f_\mathcal{S}$ vs. $g_{\mathcal{T}->\mathcal{S}}$. This comparison demonstrates if training bias mitigation networks can benefit from external knowledge from teacher models, which answers *RQ2* (§4.2).

We note that when $\mathcal{T} = \mathcal{S}$, C1 and C2 are essentially the same comparison. To avoid duplicate discussion, we will present results when $\mathcal{T} = \mathcal{S}$ in C2.

3 Experimental Setup

For consistency and fair comparison with previous debiasing works in NLU (Jeon et al., 2023; Reif and Schwartz, 2023) and image classification (Kirichenko et al., 2023; LaBonte et al., 2023; Li et al., 2023), we adopt commonly used experimental setups, including choice of backbone models, datasets, evaluation protocols, and debiasing methods. In addition, all experiments are repeated three times with different random seeds to account for any stochastic effect.

Backbones We conduct experiments on a series of BERT (Devlin et al., 2019; Turc et al., 2019), T5 (Tay et al., 2022), ResNet (He et al., 2016), and ViT (Dosovitskiy et al., 2021) backbones of different scales, shown in Appendix D Table 2. These backbones are chosen for several reasons: BERT and ResNet are commonly employed in prior works, which enables consistent comparisons. In addition, ViT and T5 are commonly used backbones for vision and language tasks, but relatively less experimented in prior debiasing works, which allows investigating the generalizability of our findings beyond existing research. Finally, each backbone is associated with a series of publicly available pre-trained checkpoints of different scales, with consistent network architecture and pre-training data, which enables cross-scale distillation and comparisons.

Evaluation To provide a comprehensive evaluation of robustness against spurious correlations,

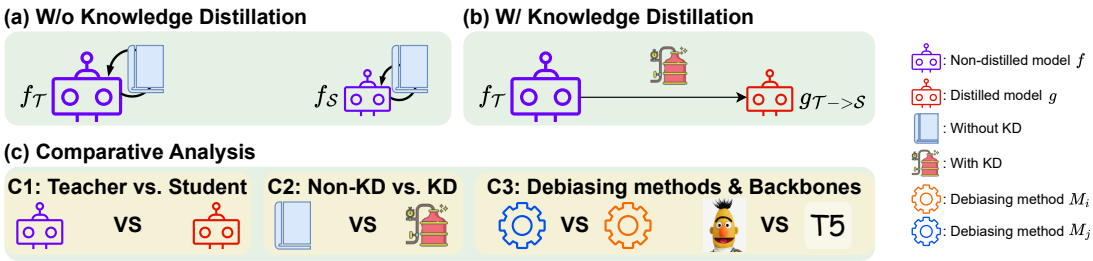


Figure 1: Framework for the analysis of distillability of debiasing methods. **(a) training from scratch**: we train a debiasing method M_i from scratch without knowledge distillation on different scales (teacher $\mathcal{T}$ and student $\mathcal{S}$ such that $\mathcal{T} > \mathcal{S}$) to obtain models $f_\mathcal{T}$, $f_\mathcal{S}$ respectively. **(b) Training with knowledge distillation**: we apply knowledge distillation to transfer knowledge from teacher ($f_\mathcal{T}$) to student ($g_{\mathcal{T}->\mathcal{S}}$). **(c) Assessment**: *C1* determines if knowledge distillation can transfer the debiasing capability from teacher ($f_\mathcal{T}$) to student ($g_{\mathcal{T}->\mathcal{S}}$), *C2* determines the contribution of knowledge distillation in training a debiased model, and *C3* compares different debiasing methods and backbones under knowledge distillation.

we compare the teacher $f_\mathcal{T}$ and the student $g_{\mathcal{T}->\mathcal{S}}$ from the following perspectives:

- **In-domain performance (ID, ↑)**: the performance on in-domain test set. A robust model should achieve high performance on this set to demonstrate general capability.
- **Out-of-domain performance (OOD, ↑)**: the performance on in-domain test set. Such samples require real task-related signals to predict, where biased models fall short. For text datasets, we evaluate models on separate OOD test sets, comprised with specially crafted hard samples (McCoy et al., 2019). For image datasets, samples are divided into groups based on their labels and spurious attributes, where OOD refers to the worst performing sub-group (Yang et al., 2023).
- **Spurious gap (Spu. Gap, ↓)**: the performance gap between ID and OOD, which quantifies a model's vulnerability to spurious correlations. Ideally, a robust model should have high performances on both ID and OOD with a small spurious gap.

Similarly, we compare KD and Non-KD as above. We compute F1 score on QQP and accuracy on other datasets.

Investigating Internal Mechanism Besides the above superficial performance metrics, we aim to uncover the internal mechanisms that causes the change of debiasing ability post-KD.

- **Activation Pattern**. We conduct activation-level analysis and comparisons across layers. We use Centralized Kernel Alignment (CKA), a commonly adopted technique to measure the similarity between activation matrices or hidden representations of neural networks (Kornblith et al., 2019; Cortes et al., 2012). Following previous work (Raghu et al., 2021; Nguyen et al., 2021), we use CKA by first probing the intermediate representations from each layer and then comparing all pairwise similarities between representations of the teacher and student models, under linear kernel CKA.
- **Circuit Discovery**. We also analyze and compare bias-specific sub-networks, or "circuits". This approach moves beyond simply observing a model's outputs to causally trace how information flows through and is processed by a coordinated set of components. Specifically, we use EAP (Hanna et al., 2024), a widely adopted method for circuit discovery.

Datasets We use the following datasets: 1) CelebA (Liu et al., 2015), 2) Waterbird (Sagawa et al., 2020), 3) MNLI (Williams et al., 2018), and 4) QQP (Sharma et al., 2019). More details of dataset statistics, causal and spurious features, and the OOD test sets are discussed in Appendix B.

Debiasing Methods Experiments are conducted on a comprehensive list of commonly used debiasing methods, each of which is designed with special formulation and assumptions. We use (a) Empirical Risk Minimization (ERM) (standard training without debiasing techniques, (b) HypothesisOnly-PoE (Karimi Mahabadi et al., 2020), (c) WeakLearner-PoE (Sanh et al., 2021), (d) KernelWhitening (Gao et al., 2022), (e) AttentionPoE (Wang et al., 2023), (f) CurriculumDebi-

asing (Lee et al., 2025), (g) σ-Damp (Puli et al., 2023), (h) DeepFeatReweight (Kirichenko et al., 2023), and (i) PerSampleGrad (Ahn et al., 2023). The above debiasing methods have a wide coverage of existing algorithms, ranging from auxiliary biased model-based debiasing, to disentanglement of representations. Meanwhile, they can handle multiple types of shortcuts at the same time, without overfitting to a specific bias. Details of these methods are provided in Appendix C.

4 Effect of Knowledge Distillation on Debiasing

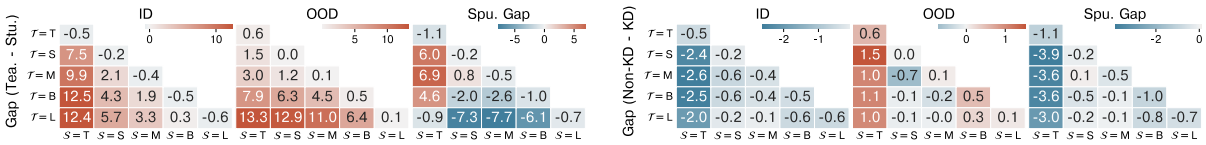
We first examine if KD can effectively distill the debiasing capability from teachers to students of different scales in Section 4.1. We then asses if training with knowledge distillation (KD) can improve a model's debiasing performance compared to standard training (Non-KD) in Section 4.2. Finally, we assess the effect of different debiasing methods and backbones on our earlier findings in Section 4.3.

4.1 RQ1: Distillability of Debiasing Methods

Students become more biased than teachers We observe that teachers consistently achieve better performance than their smaller scale students on ID and OOD test sets after knowledge distillation. The positive values in Figure 2a show that although KD encourages students to mimic their teachers in the logit space, it may undesirably increase student's susceptibility to spurious correlations in datasets as the teacher's in-domain and debiasing capabilities are not effectively transferred to the student. The prediction agreement between teacher and student models show similar trend, where the student generally aligns with the teacher on ID but often largely diverges from the teacher on OOD. Furthermore, the extent of knowledge loss after distillation varies depending on the relative scales of the teacher and student models. For example, as depicted in Figure 2a, when $\mathcal{S}$ is tiny ($\mathcal{S}$ = T), more debiasing power is lost, shown by mostly positive values in spurious gap. When $\mathcal{T}$ is large ($\mathcal{T}$ = L), more ID knowledge is lost, shown by mostly negative values in spurious gap. The results show that if a teacher model learns a partially debiased representation but still retains residual biases, the student might amplify this bias rather than mitigate it.

Students show diverse distribution shifts in predicted probabilities To understand the influence of KD on the debiasing capabilities of students, we investigate the output probability distribution $P_{\mathcal{C}}(y = 1)$. Our findings show that KD significantly alter the predicted probability distribution, despite its training objective of matching output logits. This perturbation is often larger on OOD than ID test sets, which explains the larger performance drop observed in students compared to their teachers on OOD, as illustrated in Figure **??**. We also observed that teachers tend to provide slightly more confident predictions on ID while more moderate predictions on OOD. Such behavior is not successfully transferred to students through KD. Such distinct behaviors on different samples may encourage models to overfit to data distributions of the training sets or to over-represented groups, which can effectively amplify reliance on shortcuts over robust features. In addition, the training sets of teachers often contain biased examples or do not equally represent all sub-groups, which leads students to inherit and potentially amplify these biases. Consequently, students often perform worse than their teachers on OOD.

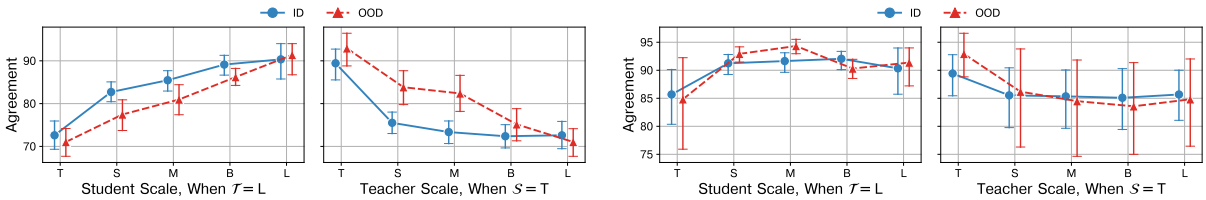
Potential for new debiasing capabilities for students beyond teacher abilities We compare prediction agreement between teacher and student models. When $\mathcal{T}$ is large ($\mathcal{T}$ = L), we observe an increase in prediction agreement as $\mathcal{S}$ scales up, with consistently higher agreement on ID than OOD, as shown in the left plot in Figure 3. Conversely, when $\mathcal{S}$ is tiny ($\mathcal{S}$ = T), the prediction agreement diminishes as $\mathcal{T}$ scales up, with higher agreement on OOD than ID, the right plot in Figure 3. The imperfect agreement between teacher and student contradicts with the foundational assumptions of knowledge distillation, which assumes that students should closely mimic their teachers. However, interestingly, this unexpected behavior may not always lead to performance degradation. Sometimes it enables students to generalize to out of domain data. In particular, there are instances where students make correct predictions where their teachers do not, see the left plot in Figure 5. Students can sometimes outperform their teachers perhaps because they may learn additional patterns during the knowledge distillation process, which allows them to generalize better than their teachers. The above result suggests that students may sometimes acquire debiasing capabilities that surpass those of their teachers, which we believe is a novel avenue for robust model training.



(a) C1: Teacher vs. Student

(b) C2: Non-KD vs. KD

Figure 2: **Average performance gaps** on ID, OOD, and Spurious Gap between (a) Teacher and Student and (b) Non-KD and KD. X-axis and Y-axis show the scale of student ($\mathcal{S}$) and teacher ($\mathcal{T}$) respectively. Each cell shows the performance gap. See Appendix F for detailed results.



(a) C1: Teacher vs. Student

(b) C2: Non-KD vs. KD

Figure 3: **Prediction agreement on text datasets**. Agreement increases as the scale of teacher and student get closer. See Appendix F for detailed results.

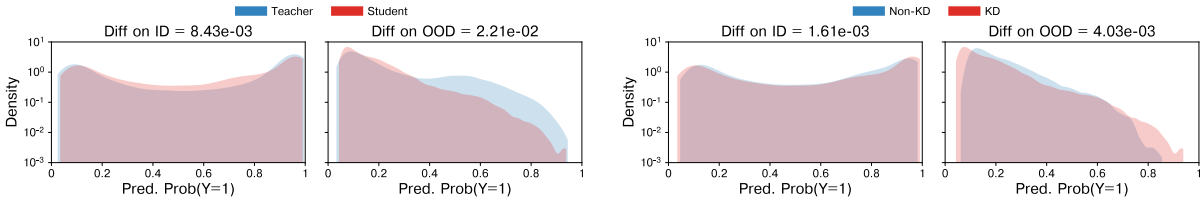
Larger teachers do not guarantee more robust students Our findings show that a more capable teacher does not guarantee a less biased student in debiasing tasks. With a fixed student scale (as seen in the columns of Figure 2), increasing the teacher's scale does not consistently reduce performance gap or spurious gap. Sometimes, a larger teacher may degrade the debiasing capability of the student. For example, when $\mathcal{S}$ = T, increasing the teacher scale from M to B increases the spurious gap from 6.5 to 8.1 on ERM, i.e. a more biased model. Moreover, when when $\mathcal{S}$ = T, increasing $\mathcal{T}$ result in a drop of teacher-student agreement, indicating that the student fails to follow the teacher, see right plot in Figure 3. We attribute this finding to two reasons. Firstly, the capabilities of students are substantially bounded by their scale, and using a much larger and capable teacher may exceed the student's capacity for effective learning (Cho and Hariharan, 2019). Secondly, training students with debiasing objectives and knowledge distillation at the same time results in optimization problem, which may trap students' parameters in local optima and affect their robustness to spurious correlations.

Students with similar scales to their teachers learn better The effectiveness of debiasing ability transfer through distillation is greatly affected by the scale similarity between teacher and student. As the teacher and the student become similar in scale (near the diagonal cells in Figure 2), the differences on test set performance and spurious gaps decrease. However, a larger mismatch in scale (far from diagonal) results in more pronounced differences, see Figure 2. Similarly, the teacher-student agreement increases as $\mathcal{T}$ and $\mathcal{S}$ align more closely, see Figure 3. This is likely because models of similar scales have comparable expressive power and extracts similar features, which can lead to more effective knowledge transfer, better bias mitigation, and higher prediction agreement.

4.2 RQ2: Distillation vs. Standard Training

Non-KD is less biased than KD Our results show that debiasing models trained from scratch (Non-KD) have lower ID performance than those trained with KD. However, the Non-KD models achieve almost no changes on OOD, leading to smaller spurious gaps, see Figure 2. We hypothesize that the distillation objective of matching logits, despite effective on ID, may potentially inject additional spurious correlations and distract the model from prioritizing robust features, as the teacher is not fully unbiased.

KD does not improve generalization An interesting finding is that both Non-KD and KD have similar average prediction agreements on both ID and OOD. However, the agreement on OOD varies significantly depending on dataset, debiasing method, and backbone model. This suggests that training solely with the original data (Non-KD) is sufficiently effective for debiasing, and intro-



(a) C1: Teacher vs. Student

(b) C2: Non-KD vs. KD

Figure 4: **Density of predicted probability**. On OOD, students has larger deviation in prediction confidence than teachers. See Appendix F for detailed results.

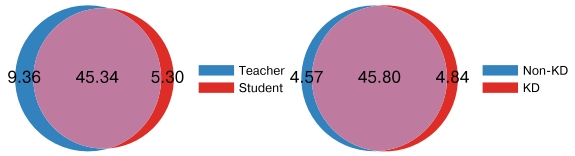


Figure 5: **C1: Teacher vs. Student** (Left) and **C2: Non-KD vs. KD** (Right): correctly predicted examples on OOD on text datasets.

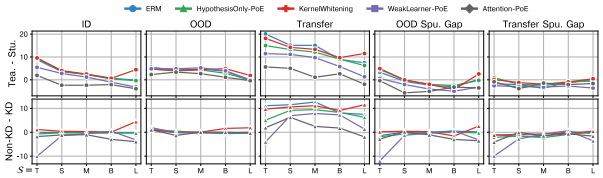


Figure 6: **Comparison Between Debiasing Methods**: performance gap between Teacher and Student (Above), Non-KD and KD (Lower) on text datasets. Detailed results are shown in Appendix.

ducing external knowledge via KD does not yield significant improvements. This result can be attributed to KD's impact on model confidence; models trained with KD tend to produce more confident predictions than models trained without KD, see Figure 4, which is key to degenerate performance on OOD (Utama et al., 2020; Sanh et al., 2021). Such overconfidence could be a critical factor in degraded performance on OOD tasks. Moreover, such minimal contribution of KD remains unchanged even when stronger external knowledge (a larger teacher) or a more capable learner (a larger student) is used, see Figure 3.

4.3 Distillability Across Methods and Backbones

Different transfer patterns across methods Our results show that the transfer patterns are heavily influenced by the formulation of debiasing method. For example, logit-based PoE methods, such as HypothesisOnly-PoE and WeakLearner-PoE, show similar trends in performance changes and spurious gaps, in contrast to the representation disentanglement method (KernelWhitening), see Figure 6.

Sensitivity to backbones The distillability of KD appears to varies with the architecture of the backbones and randomness in the training. KernelWhitening and WeakLearner-PoE are two methods particularly sensitive to the scale of backbone and random seeds, which controls factors such as data sampling and ordering.

Robustness to different biases transfer differently We observe that OOD and Transfer show different transfer patterns, where performance gap on Transfer exhibit much larger variations the student scales up, see Figure 6. This suggests that smaller students may outperform larger ones on OOD, indicating that during KD, larger students may become more prone to certain biases (OOD) but more resilient to others.

Universal transfer patterns in debiasing methods A number of debiasing methods show consistent changes in robustness after KD, which suggest the potential for an empirical universal transfer pattern. Specifically on text datasets, the performance gap between teacher and student models on OOD and Transfer Spurious Gap fall in the range of [0, 5] and [-5, 0] respectively, see Figure 6. Such change in performance is consistent across different scales of $\mathcal{T}$ and $\mathcal{S}$, which allows for predictable performance after KD. Similarly, the performance gap between models trained using Non-KD and KD remains stable on OOD, falling in range [-1, 1] across different scales.

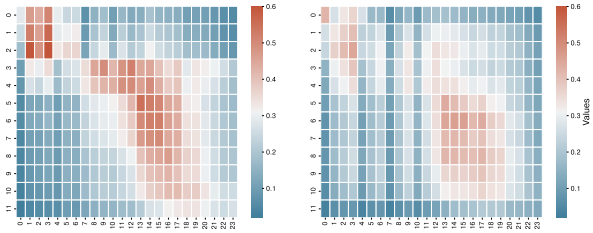


Figure 7: **C1: Teacher vs. Student**: Centered Kernel Alignment on ID (left) and OOD (right). Highers values indicate higher similarity. X-axis and Y-axis refer to the layers of teacher ($\mathcal{T}$) and student ($\mathcal{S}$) respectively. See Appendix F for detailed results.

5 Internal Mechanism of Lack of Distillability

5.1 Attention

Attention plays a critical role in making predictions, and biased models may learn spurious attention patterns (Wang et al., 2023). We hypothesize that the divergence of teacher and student on OOD samples is due to the difference in their internal attention patterns.

Figure 7 shows the difference between internal representations between teacher and student when making predictions on ID and OOD data. After distillation, students try to mimic teachers on ID (left). The earlier layers of students follow earlier layers of teachers, and similarly mid and later layers. This indicates that KD can transfer knowledge of ID data from the larger teachers into smaller students. On OOD (right), however, we observe similar pattern but it is not fully preserved. In particular, it is challenging for the mid and later layers of the students to follow closely to those of the teachers, which explains the performance degradation on OOD after KD, see Figure 7.

5.2 Circuit

Teacher vs. Student For teacher models, we observe a consistently moderate positive effect across all attention heads. While the effect of MLP layers can be negative. However, knowledge distillation have reverted the effect of attention heads into negative. There is not a unified pattern on the MLP layers.

Non-KD vs. KD On Non-KD trained models, attention heads across all layers have negative effect on the final logit, except for the final MLP layer which has a strong positive effect. On the contrary, the KD trained models tend to emphasize MLP layers and suppress the contribution on attention heads. This is likely due to learning process of KD is centered on matching logits, which may miss some contributions of the earlier layers.

6 Potential Solutions

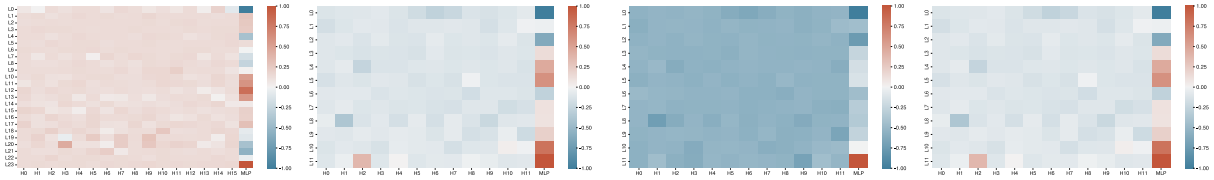
Based on the above analyses, we summarize the key findings on distilling debiasing models as follows:

- Training distribution significantly affect successful distillation of debiasing capabilities.
- Student models with similar scale to their teachers can better obtain debiasing knowledge from their teachers.
- The objectives of KD may introduce additional optimization challenges, especially with the presence of debiasing objectives.

To further improve the distillability of debiasing methods, we propose three solutions:

Data augmentation (DA) There is broad evidence that models becomes biased by relying on spurious features in the training set (Wu et al., 2022; Ahn et al., 2023), which is amplified by misrepresentation of specific classes or labeling errors. Prior studies have highlighted the important role of data in knowledge distillation (Stanton et al., 2021). Based on these prior studies and our findings, we hypothesize that providing high quality data and augmenting data size can improve the process of distilling the debiasing capability from teacher to student. For text datasets, we employ the data generated by Seq-Z filtering (Wu et al., 2022) as training set for both teacher and student models. For image datasets, we employ training and validation sets where the sub-groups are equally represented (Kirichenko et al., 2023).

Iterative knowledge distillation (IKD) Our results indicate that the transfer of debiasing capabilities is more effective between teachers and students of similar scales. Therefore, we propose to leverage Iterative Knowledge Distillation (IKD) (Liu et al., 2023): given a teacher of scale $\mathcal{S}_N$, we first distill it to a student of scale $\mathcal{S}_{N-1}$, where $\mathcal{S}_{N-1}$ is the closest neighbor of $\mathcal{S}_N$ in scale. Then the newly-distilled student acts as a teacher and transfer the knowledge to a model of smaller scale $\mathcal{S}_{N-2}$, where $\mathcal{S}_{N-2}$ is the closest neighbor of $\mathcal{S}_{N-1}$ in scale. We repeat this process iteratively by gradually decreasing student scale, such that the knowledge can be transferred smoothly from a large scale



(a) C1: Teacher (Left) vs. Student (Right)

(b) C2: Non-KD (Left) vs. KD (Right)

Figure 8: **Comparison of Discovered Circuit**. L, H, MLP denotes the layer, attention head and MLP components respectively. Each cell shows the causal effect on logit.

model to a small scale model. This step-wise approach enables a smooth knowledge transfer from larger to smaller scale models, and potentially improves debiasing effectiveness at each step.

Initialize student with teacher weights (Init) Previous research by Stanton et al. (2021) has discovered that initializing a student model with the weights of its teacher can increase their centered kernel alignment (Kornblith et al., 2019) in activation space. This approach can head-start the student model with a stronger debiasing capability from the teacher. It can also help alleviate potential optimization obstacles and stuck in local optima. If the teacher and student models are of the same scale, we initialize the student with the teacher parameters. If the teacher is larger, we initialize the student with the first few layers of the teacher.

Results Table 1 shows that all three solutions result in improved distillability. We additionally compared to robustness-preserving KD methods, Domain-Targeted Augmentation (DTA) (Stacey and Rei, 2024) and Robust Model Compression (RMC) (). Specifically, on spurious gap between teacher and student, data augmentation (DA), iterative knowledge distillation (IKD), and initialization with teacher weights (Init) yield performance gains of 4.5, 2.8, 1.2 absolute points compared to vanilla KD across datasets and backbones respectively. On spurious gap between Non-KD and KD, DA, IKD, and Init outperforms vanilla KD by 1.7, 0.6 and 0.2 absolute points respectively. We find that DA achieves the largest improvement, since the root cause of spurious correlations come from the underlying dataset (Chen et al., 2018). Debiasing the dataset itself can benefit all training methods including knowledge distillation. As noted by previous work (Stanton et al., 2021), Init may facilitate teacher-student agreement in activation space, but result in non-significant gains, which aligns with our findings as well.

Table 1: Improvement of distillability. Vanilla refers to standard knowledge distillation, DA, IKD, and Init represent data augmentation, iterative knowledge distillation, and initialization of student with teacher weights (Init) as our three solutions to improve the distillability of debiasing methods.

Difference in	ID (↓)	OOD (↓)	Spu. Gap (↓)
Teacher - Student			
Vanilla	5.1	7.3	12.7
+ DA	2.3	5.4	8.2
+ IKD	3.6	5.9	9.9
+ Init.	4.7	6.5	11.5
Non KD - KD			
Vanilla	1.4	0.7	2.2
+ DA	0.2	0.2	0.5
+ IKD	1.0	0.5	1.6
+ Init.	1.3	0.7	2.0

7 Conclusion

We present the first study on the distillability of debiasing capabilities between neural models, and the extent of bias transfer through knowledge distillation (KD). We evaluate eight popular debiasing methods and five scales of backbones on four datasets. Extensive experiments show that vanilla KD does not consistently preserve debiasing capabilities; in many cases, student models become more reliant on spurious correlations than their teachers; the effectiveness of debiasing transfer depends on model scale similarity–distillation works best when teacher and student models are comparable in complexity; and larger teachers do not always yield more robust students, which indicates the need for targeted debiasing strategies in KD. We propose three solutions–data augmentation, iterative KD, and student initialization–which significantly improve the distillability of debiasing methods and contribution of KD on debiasing.

In future we will investigate self-distilled debiasing, where the student iteratively distills knowledge from itself rather than relying on a fixed teacher. A potential improvement is to explicitly guide the student using counterfactual data augmentation during distillation.

Limitations

Despite delivering significant amount of discoveries, our work has certain limitations. Firstly, our experiments are mainly conducted on logit-based knowledge distillation. The effect of other knowledge distillation methods has not been explored. Secondly, the work does not explore the scenario where multiple teachers participate in the distillation process.

Ethical Considerations

Our research focuses on mitigating dataset biases in text and vision datasets, and understanding why debiasing methods may fail under knowledge distillation. The broader impacts of our work are in advancing dataset fairness and potentially enhancing decision-making based on data. Our work contributes to improving the accuracy and reliability of NLP and vision models, as well as their trust and adoption.

References

Sumyeong Ahn, Seongyoon Kim, and Se-Young Yun. 2023. Mitigating dataset bias by using per-sample gradient. In *The Eleventh International Conference on Learning Representations*.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. 2020. Learning de-biased representations with biased representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 528–539. PMLR.

Simone Bombari and Marco Mondelli. 2025. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and overparameterization. In *Forty-second International Conference on Machine Learning*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. 2021. Cross-layer distillation with semantic calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7028–7036.

Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Ziheng Chen, Jiali Cheng, Hadi Amiri, Kaushiki Nag, Lu Lin, Xiangguo Sun, and Gabriele Tolomei. 2025. Frog: Fair removal on graphs. *arXiv preprint arXiv:2503.18197*.

Jiali Cheng and Hadi Amiri. 2024a. FairFlow: Mitigating dataset biases through undecided learning for natural language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975, Miami, Florida, USA. Association for Computational Linguistics.

Jiali Cheng and Hadi Amiri. 2024b. Mu-bench: A multitask multimodal benchmark for machine unlearning. *arXiv preprint arXiv:2406.14796*.

Jiali Cheng and Hadi Amiri. 2025. EqualizeIR: Mitigating linguistic biases in retrieval models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 889–898, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiali Cheng, Mohamed Elgaar, Nidhi Vakil, and Hadi Amiri. 2024. CogniVoice: Multimodal and Multilingual Fusion Networks for Mild Cognitive Impairment Assessment from Spontaneous Speech. In *Interspeech 2024*, pages 4308–4312.

Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. 2024. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1013–1025, St. Julian's, Malta. Association for Computational Linguistics.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. 2012. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(28):795–828.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

SongYang Gao, Shihan Dou, Qi Zhang, and Xuanjing Huang. 2022. Kernel-whitening: Overcome dataset bias with isotropic sentence embedding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4112–4122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qi Guo, Yuanhang Tang, Yawen Ouyang, Zhen Wu, and Xinyu Dai. 2023. Debias NLU datasets via training-free perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10886–10901, Singapore. Association for Computational Linguistics.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*.

Changyi He, Yifu Ding, Jinyang Guo, Ruihao Gong, Haotong Qin, and Xianglong Liu. 2025. DA-KD: Difficulty-aware knowledge distillation for efficient large language models. In *Forty-second International Conference on Machine Learning*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. 2019. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3779–3787.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800.

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

Eojin Jeon, Mingyu Lee, Juhyeong Park, Yeachan Kim, Wing-Lam Mok, and SangKeun Lee. 2023. Improving bias mitigation through bias experts in natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11053–11066, Singapore. Association for Computational Linguistics.

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.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Nayeong Kim, SEHYUN HWANG, Sungsoo Ahn, Jaesik Park, and Suha Kwak. 2022. Learning debiased classifier with biased committee. In *Advances in Neural Information Processing Systems*, volume 35, pages 18403–18415. Curran Associates, Inc.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. 2023. Towards last-layer retraining for group robustness with fewer annotations. In *Advances in Neural Information Processing Systems*, volume 36, pages 11552–11579. Curran Associates, Inc.

Mingyu Lee, Yeachan Kim, Wing-Lam Mok, and SangKeun Lee. 2025. Curriculum debiasing: Toward robust parameter-efficient fine-tuning against dataset biases. In *Proceedings of the 63rd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9524–9540, Vienna, Austria. Association for Computational Linguistics.

Weiwei Li, Junzhuo Liu, Yuanyuan Ren, Yuchen Zheng, Yahao Liu, and Wen Li. 2025. Let samples speak: Mitigating spurious correlation by exploiting the clusterness of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15486–15496.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. 2023. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20071–20082.

Jiajun Liu, Peng Wang, Ziyu Shang, and Chenxiao Wu. 2023. Iterde: An iterative knowledge distillation framework for knowledge graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:4488–4496.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738.

Yougang Lyu, Piji Li, Yechang Yang, Maarten de Rijke, Pengjie Ren, Yukun Zhao, Dawei Yin, and Zhaochun Ren. 2022. Feature-level debiased natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in NLU. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7607–7613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pages 20673–20684. Curran Associates, Inc.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. 2021. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*.

Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. 2024. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27662–27671.

Aahlad Manas Puli, Lily Zhang, Yoav Wald, and Rajesh Ranganath. 2023. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. In *Advances in Neural Information Processing Systems*, volume 36, pages 71874–71910. Curran Associates, Inc.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128. Curran Associates, Inc.

Abhilasha Ravichander, Joe Stacey, and Marek Rei. 2023. When and why does bias mitigation work? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9233–9247, Singapore. Association for Computational Linguistics.

Yuval Reif and Roy Schwartz. 2023. Fighting bias with bias: Promoting model robustness by amplifying dataset biases. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13169–13189, Toronto, Canada. Association for Computational Linguistics.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *arXiv e-prints*.

Joe Stacey and Marek Rei. 2024. Distilling robustness into natural language inference models with domain-targeted augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2239–2258, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Don Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew Gordon Wilson. 2021. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*.

Md Sultan. 2023. Knowledge distillation ≈ label smoothing: Fact or fallacy? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4469–4477, Singapore. Association for Computational Linguistics.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. 2024. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15731–15740.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations*.

Rishabh Tiwari, Durga Sivasubramanian, Anmol Mekala, Ganesh Ramakrishnan, and Pradeep Shenoy. 2024. Using early readouts to mediate featural bias in distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2638–2647.

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

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. Caltech-ucsd birds-200-2011 (cub-200-2011). Technical report, California Institute of Technology.

Fei Wang, James Y. Huang, Tianyi Yan, Wenxuan Zhou, and Muhao Chen. 2023. Robust natural language understanding with residual attention debiasing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 504–519, Toronto, Canada. Association for Computational Linguistics.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3091–3100.

Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. 2020. Exclusivity-consistency regularized knowledge distillation for face recognition. In *Computer Vision – ECCV 2020*, pages 325–342, Cham. Springer International Publishing.

Yining Wang, Junjie Sun, Chenyue Wang, Mi Zhang, and Min Yang. 2024. Navigate beyond shortcuts: Debiased learning through the lens of neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12322–12331.

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.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Zenglin Xu, Rong Jin, Bin Shen, and Shenghuo Zhu. 2015. Nystrom approximation for sparse kernel methods: Theoretical analysis and empirical evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2023. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *The Eleventh International Conference on Learning Representations*.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. 2023. Change is hard: A closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39584–39622. PMLR.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models

by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada. Association for Computational Linguistics.

Sergey Zagoruyko and Nikos Komodakis. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

A Related Work

Bias mitigation in NLU Debiasing approaches usually employ a biased model to inform the training of a robust model (Clark et al., 2019; Karimi Mahabadi et al., 2020; Sanh et al., 2021; Utama et al., 2020; Cheng et al., 2024). Other methods aim at learning debiased or robust representations (Gao et al., 2022; Lyu et al., 2022; Wang et al., 2023; Jeon et al., 2023; Reif and Schwartz, 2023), or removing bias-encoding parameters (Meissner et al., 2022; Yu et al., 2023; Cheng and Amiri, 2025). Other works include measurement of bias of specific words with statistical test (Gardner et al., 2021), generating non-biased samples (Wu et al., 2022), identification of bias-encoding parameters (Yu et al., 2023), when bias mitigation works (Ravichander et al., 2023), bias transfer from other models (Jin et al., 2021), and bias removal with model unlearning (Cheng and Amiri, 2024b; Chen et al., 2025).

Bias mitigation in vision In vision, worst-group performance is measured as a sign of model robustness. Several works investigates how to learn debiased models from failure cases (Nam et al., 2020), biased representations (Bahng et al., 2020), multiple biased models (Kim et al., 2022), and by simply re-training the last layer of a neural model (i.e. the classification layer) with additional equally represented data (Kirichenko et al., 2023; LaBonte et al., 2023). Li et al. (2023) showed that multiple spurious features can occur in a dataset, while suppressing one may inevitably boost another one. Other perspectives for debiasing include causal attention (Wang et al., 2021), building uniform margin classifiers (Puli et al., 2023), using representations from earlier layers (Tiwari et al., 2024), and neural collapse (Wang et al., 2024), where feature space collapses into a stable geometric structure that results in robustness and generalizability.

Knowledge distillation Knowledge Distillation (KD) is initially proposed to transfer knowledge from a larger model (teacher) to a smaller model (student), by encouraging the student to follow the teacher on prediction logits (Hinton et al., 2015), learned features (Romero et al., 2015; Wang et al., 2020), attention map (Zagoruyko and Komodakis, 2017; Chen et al., 2021), activation patterns (Huang and Wang, 2017; Heo et al., 2019). Later works discovered that KD can be viewed as a special form of regularization similar to label smoothing (Szegedy

et al., 2016), providing no task-specific knowledge. However, on text classification tasks, whether KD can regularize the student depends on the choice of teacher model (Sultan, 2023), which may result in opposite model confidence between teacher and student compared to label smoothing. Stanton et al. (2021) discovers that optimization and dataset details are crucial to matching students to teachers, and such matching does not guarantee better generalization ability of students. Xue et al. (2023) investigates cross-modal KD, where the teacher functions on a different modality or extra modalities than student. The authors proposes modality fusing hypothesis, which claims that modality decisive features are critical for the effectiveness of cross-modal KD. However, despite briefly discussed (Cho and Hariharan, 2019; Tiwari et al., 2024), the potential of knowledge distillation to transfer debiasing capabilities across different modalities and backbone models remains under-explored and poorly understood in existing work.

B Details of Dataset

We describe the details of each dataset below:

- **CelebA (Liu et al., 2015)** consists of 16k images of celebrity faces, where the objective is to predict "Blond_Hair" given "Male" as a spurious attribution.

- **Waterbird (Sagawa et al., 2020)** consists of synthetic images of birds from CUB dataset (Wah et al., 2011) and backgrounds (land & water) from Places (Zhou et al., 2018) dataset. The objective is to correctly infer "land bird" or "water bird," given the background as misleading information.

- **MNLI (Williams et al., 2018)** consists of 39k natural language inference (NLI) samples from various domains, where the objective is to classify relationship between a premise and a hypothesis as "Entailment", "Contradiction", or "Neutral". Previous studies discover that models are prone to negation words, lexical overlap, and sub-sequence biases in NLI task (Naik et al., 2018; Mendelson and Belinkov, 2021). We use HANS (McCoy et al., 2019) as the out-of-distribution test set (OOD) and SNLI (Bowman et al., 2015) as the transfer test set (Transfer), detailed below.

- **QQP (Sharma et al., 2019)** is a paraphrase identification (PI) dataset with 43k samples, where the objective is to predict if two questions are paraphrases of each other. Similar to MNLI, models are likely to be mislead by lexical overlap between two questions. We exploit PAWS (Zhang et al., 2019) as the out-of-distribution test set (OOD) and MRPC (Dolan and Brockett, 2005) as the transfer test set (Transfer), detailed below.

C Details on Debiasing Methods

Experiments are conducted on a comprehensive list of commonly used debiasing methods, each of which is designed with special formulation and assumptions.

- **Empirical Risk Minimization (ERM)** is the standard training method that minimizes the empirical risk on a dataset. This is akin to fine-tuning a pre-trained model on a dataset using cross-entropy loss with no debiasing strategy, which works for both image and text datasets.

- **HypothesisOnly-PoE** (Karimi Mahabadi et al., 2020) assumes the hypothesis part of NLI datasets contains biases. It trains a hypothesis-only (biased) model to measure the bias of each sample, and uses Product-of-Experts (PoE) (Hinton, 2002) to adjust the confidence of the debiased model according to the confidence of the biased model. This approach is evaluated on text datasets.

- **WeakLearner-PoE** (Sanh et al., 2021) leverages weak learners to capture and model bias, including bias of unknown type. It trains a 2-layer BERT as a biased model and exploits PoE to train the debiased model. This approach is evaluated on text datasets.

- **KernelWhitening** (Gao et al., 2022) aims at learning isotropic sentence embeddings with disentangled robust and spurious representations, with Nyström kernel (Xu et al., 2015). This approach is evaluated on text datasets.

- **AttentionPoE** (Wang et al., 2023) assumes that the attention to `[CLS]` token in text classification is biased and introduces PoE on attention weights to learn robust attention patterns for bias mitigation. This approach is evaluated on text datasets.

- **σ-Damp** (Puli et al., 2023) assuming the standard cross-entropy loss encourages models to prioritize shortcuts over robust features, this model proposes to scale the loss by a temperature. This approach is evaluated on image datasets.

- **DeepFeatReweight** (Kirichenko et al., 2023) discovers that simply retraining the last layer of a neural model–the classification layer in supervised tasks–on top of the existing biased feature extractor is good strategy for bias mitigation. This approach is evaluated on image datasets.

- **PerSampleGrad** (Ahn et al., 2023) trains a debiased model with non-uniform sampling probability, obtained from per-sample gradient norm of a biased model. This approach is evaluated on image datasets.

D Implementation details

We follow previous debiasing works for implementation details. For text datasets, we train each debiasing method with Adam optimizer, learning rate $5e-5$, 5 epochs, both KD and Non-KD. For image datasets, we train each debiasing method with Adam optimizer, learning rate $4e-5$, 100 epochs, both KD and Non-KD. For all other hyperparameters, we follow each debiasing method's best-performing setting.

We show the details of backbone models in Table 2.

E Results on Image Datasets

On image datasets, we observe similar results on text datasets. Specifically, we see that KD fall short on distilling the debiasing capabilities. Such ability is transferred more smoothly as teacher and student get similar in scale.

Table 2: Different scales of backbones in our experiments. h and d denote number of hidden layers and size of hidden dimension respectively. T, S, M, B, L refer to Tiny, Small, Medium, Base and Large version of the backbone. See Appendix for more details.

Scale	BERT		T5		ResNet		ViT	
	h	d	h	d	h	d	h	d
T	2	128	4	256	18	512	12	192
S	4	256	8	384	34	512	12	384
M	8	512	16	512	50	2048	12	768
B	12	768	24	768	101	2048	24	1024
L	24	1024	48	1024	152	2048	32	1280

F Detailed Results on Debiasing Methods and Backbones

We present the detailed results of individual debiasing method and backbone below.

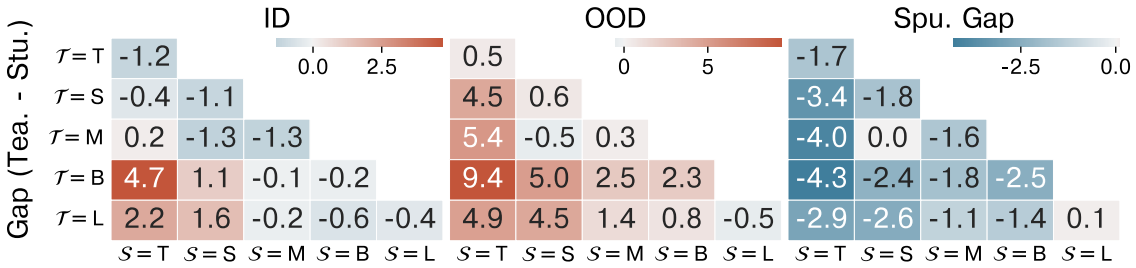


Figure 9: **C1: Teacher vs. Student**: average performance gaps between teacher and student models on ID, OOD, and Spurious Gap across image datasets. X-axis and Y-axis show the scale of student ($\mathcal{S}$) and teacher ($\mathcal{T}$) respectively. Each cell shows the performance gap between corresponding scales of a teacher and a student.

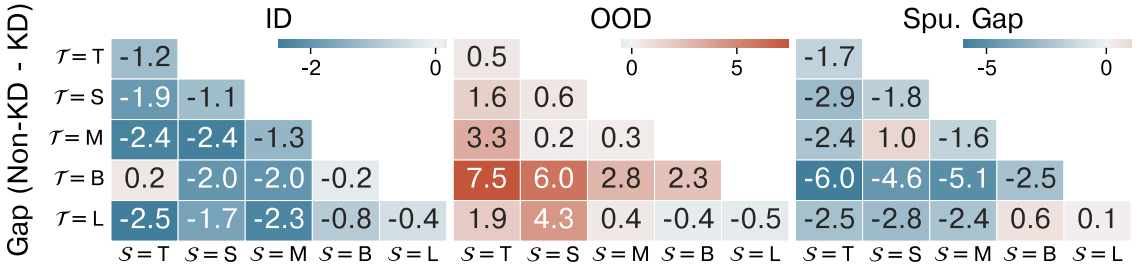


Figure 10: **C2: Non-KD vs. KD**: average performance gaps between Non-KD and KD models on ID, OOD, and Spurious Gap across image datasets. X-axis and Y-axis show the scale of student ($\mathcal{S}$) and teacher ($\mathcal{T}$) respectively. Each cell shows the performance gap between corresponding scales of a teacher and a student.

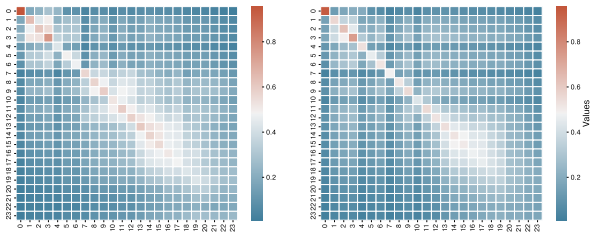


Figure 11: **C2: KD vs. Non-KD**: Centralized Kernel Alignment. Highers values indicate higher similarity. X-axis and Y-axis refer to the layers of KD ($\mathcal{S}$) and Non-KD ($f_{\mathcal{S}}$) respectively.

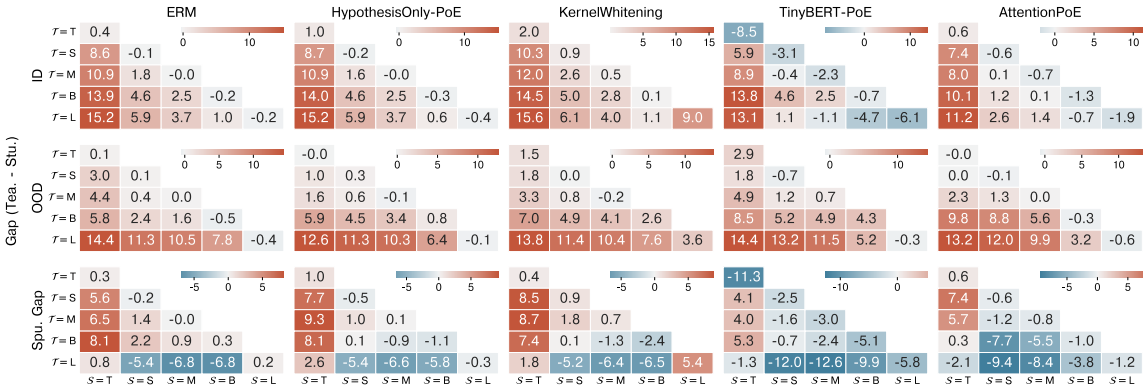


Figure 12: **C1: Teacher vs. Student**: average performance gaps between teacher and student models on ID, OOD, and Spurious Gap on BERT.

Figure 13: **C1: Teacher vs. Student**: average performance gaps between teacher and student models on ID, OOD, and Spurious Gap on T5.

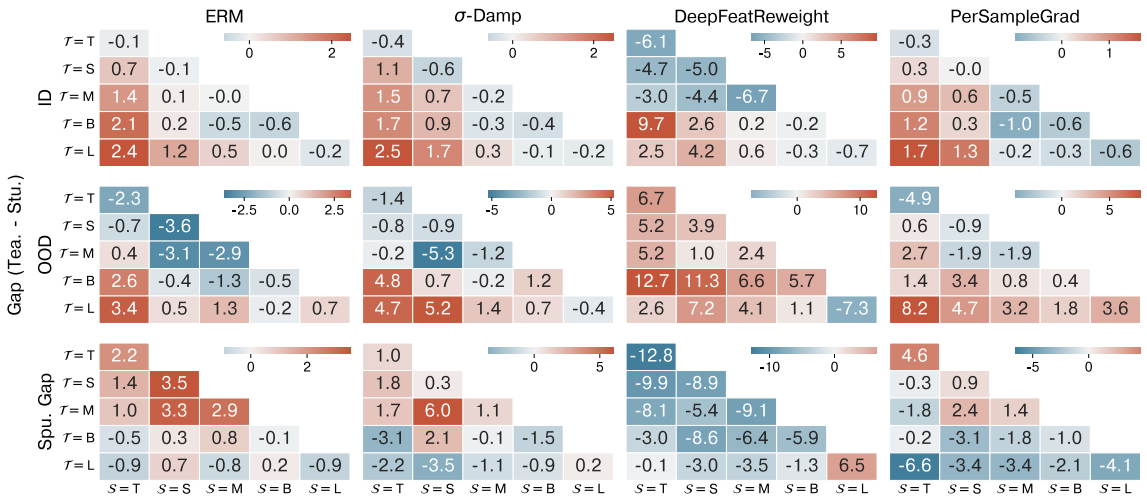


Figure 14: **C1: Teacher vs. Student**: average performance gaps between teacher and student models on ID, OOD, and Spurious Gap on ResNet.

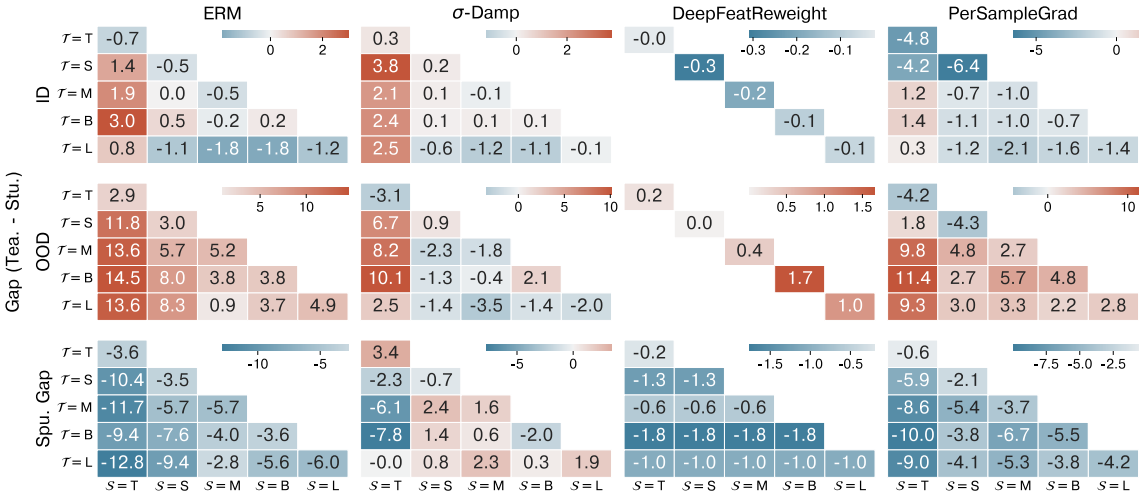


Figure 15: **C1: Teacher vs. Student**: average performance gaps between teacher and student models on ID, OOD, and Spurious Gap on ViT.

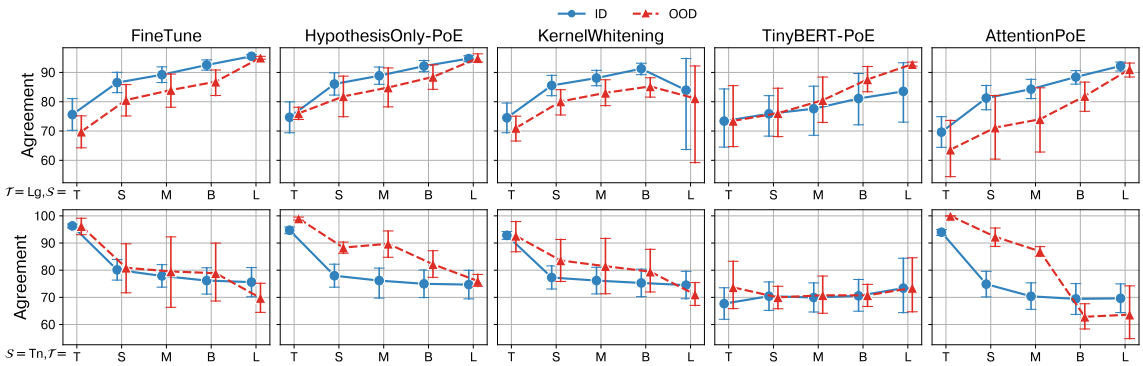


Figure 16: **C1: Teacher vs. Student**: prediction agreement on BERT. Left: varying $\mathcal{S}$ (X-axis) given a fixed teacher with $\mathcal{T}$ = L. Right: varying $\mathcal{T}$ (X-axis) given a fixed student with $\mathcal{T}$ = T. Agreement increases as the scale of teacher and student get closer.