A. Appendix

A.1. Filtering Heuristics

We present three heuristic approaches that approximate the optimum bias reduction problem (AFOPT): **(A)** A simple *greedy approach* starts with the full set $S = \mathcal{D}$, identifies an $i \in S$ that maximizes $\tilde{p}(i)$, removes it from S, and repeats up to $|\mathcal{D}| - n$ times. **(B)** A *greedy slicing approach* identifies the instances with the k highest predictability scores, removes all of them from S, and repeats the process up to $\lfloor \frac{|\mathcal{D}|-n}{k} \rfloor$ times. **(C)** A *slice sampling approach*, instead of greedily choosing the top k instances, randomly samples k instances with probabilities proportional to their predictability scores (cf. Appendix §A.2 for more details).

All three strategies could be further improved by considering not only the predictability score of the top-k instances but also (via retraining without these instances) how their removal would influence the predictability scores of other instances in the next step. We found our computationally lighter approaches to work well even without the additional overhead of such look-ahead. AFLITE implements the greedy slicing approach, and can thus be viewed as a scalable and practical approximation of (intractable) AFOPT for optimum bias reduction. We leave the empirical investigation into other proposed strategies for future work.

A.2. Slice Sampling Details

As discussed in Appendix §A.1 **(C)**, the *slice sampling approach* can be efficiently implemented using what is known as the Gumbel method or Gumbel trick (Gumbel & Lieblein, 1954; Maddison et al., 2014), which uses random perturbations to turn sampling into a simpler problem of optimization. This has recently found success in several probabilistic inference applications (Kim et al., 2016; Jang et al., 2016; Maddison et al., 2016; Balog et al., 2017; Kool et al., 2019). Starting with the log-predictability scores $\log \tilde{p}(i)$ for various i, the idea is to perturb them by adding an independent random noise γ_i drawn from the standard Gumbel distribution. Interestingly, the maximizer i^* of $\gamma_i + \log \tilde{p}(i)$ turns out to be an exact sample drawn from the (unnormalized) distribution defined by $\tilde{p}$. Note that i^* is a random variable since the γ_i are drawn at random. This result can be generalized (Vieira, 2014) for slice sampling: the k highest values of Gumbel-perturbed log-predictability scores correspond to sampling, without replacement, k items from the probability distribution defined by $\tilde{p}$. The Gumbel method is typically applied to exponentially large combinatorial spaces, where it is challenging to scale up. In our setting, however, the overhead is minimal since the cost of drawing a random γ_i is negligible compared to computing $\tilde{p}(i)$.

Table 7. Mean *Dev* accuracy (%) on two models trained on four synthetic datasets before (D) and after ($D(\Phi)$) AFLITE. Standard deviation across 10 runs with randomly chosen seeds is provided as a subscript. The datasets, also shown in Fig. 2 differ in the degree of separation between the two classes. Both models (SVM with an RBF kernel & linear classifier with logisitic regression) perform well on the original synthetic dataset, before filtering. The linear classifier performs well on the data, because it contains spurious artifacts, making the task artificially easier for it. However, after AFLITE, the linear model, relying mostly on the spurious features, clearly underperforms.

Class Separation	Model	D	$D(\Phi)$
0.8	SVM-RBF	97.0_{02}	90.7_{06}
	Logistic Reg.	83.5_{05}	50.7_{12}
0.7	SVM-RBF	89.9_{04}	82.5_{09}
	Logistic Reg.	74.3_{11}	52.4_{13}
0.6	SVM-RBF	87.6_{04}	77.8_{09}
	Logistic Reg.	74.3_{10}	53.1_{12}
0.4	SVM-RBF	83.8_{05}	70.7_{10}
	Logistic Reg.	75.4_{09}	53.4_{12}

A.3. Results on Synthetic Data Experiments

As discussed in Section §3, Figure 2 shows the effect of AFLITE on four synthetic datasets containing data arranged in concentric circles at four degrees of class separation. For greater visibility, we have provided the accuracies of the SVM with RBF kernel and logistic regression in Table 7.

In summary, a stronger model such as the SVM is more robust to the presence of artifacts than a simple linear classifier. Thus, the implications for real datasets is to move towards models designed for reasoning about a specific task, hence avoiding a dependence on spurious artifacts.

A.4. NLI Out-of-distribution Benchmarks

We describe the four out-of-distribution evaluation benchmarks for NLI from Section §4.1 below:

- HANS (McCoy et al., 2019b) contains evaluation examples designed to avoid common structural heuristics (such as word overlap) which could be used by models to correctly predict NLI inputs, without true inferential reasoning.
- NLI Diagnostics (Wang et al., 2018a) is a set of handcrafted examples designed to demonstrate model performance on several fine-grained semantic categories, such as logical reasoning and commonsense knowledge.
- Stress tests for NLI (Naik et al., 2018) are a collection of tests targeting the weaknesses of strong NLI models, to check if these are robust to semantics (competence),

Table 8. Hyperparameters for the AFLITE algorithm, used for in-distribution benchmark estimation on different datasets. m denotes the size of the support of the expectation in Eq. (4), t is the training set size for the linear classifiers, k is the size of each slice, and τ is an early-stopping filtering threshold. For ImageNet, we set $n = 640K$ and hence do not need to control for τ. In every other setting, we set τ as above, and hence do not need to control for n. Detailed definitions for each hyperparameter is provided in Section §2.

	Synthetic	SNLI	MultiNLI	QNLI	ImageNet
m	128	64	64	64	64
t	100	50K	40K	10K	32.7K
k	1	10K	10K	2K	33.6K
τ	0.75	0.75	0.75	0.75	-

irrelevance (distraction) and typos (noise).

- Adversarial NLI (Nie et al., 2019) consists of premises collected from Wikipedia and other news corpora, and human generated hypotheses, arranged at different tiers of the challenge they present to a model, using a human and model in-the-loop procedure.

Recent work (McCoy et al., 2019a) has observed large variance on out-of-distribution test sets with random seeds. Hence, we report the mean and variance across 5 random seeds in all settings in Table 1. Since Adversarial NLI involves finetuning the model, and not just reporting on a different test set, we skip this step in Table 2.

A.5. Hyperparameters for AFLITE

Table 8 shows hyperparameters used to run AFLITE to obtain filtered subsets for in-distribution benchmark estimation on different datasets. Target dataset size, n and the early stop filtering threshold τ are interdependent, as the predictability score threshold determines what examples to keep, which in turn influences the desired size of the dataset, n. For ImageNet, we set $n = 640K$ and do not control for τ. We use much larger values for t and k for ImageNet than in all NLP experiments, where the use of powerful language representations (such as *RoBERTa*) allows us to get reasonable performance even with smaller training sets; ImageNet does not offer any such benefits arising from pretrained representations.

For all out-of-distribution NLP experiments, we explicitly control for the size of n, as discussed in the corresponding sections in the paper. In these cases, we typically end up using slightly larger n, allowing for the final models to get more exposure to task data which is, to a degree, helpful for out-of-distribution generalization. In ImageNet, we use the same hyperparameters in both sets of experiments. In particular, we explicitly set $n = 182K$ for SNLI, and $n = 640K$ for ImageNet AFLITE-filtering for the out-of-distribution generalization experiments.

A.6. Hyperparameters for NLP experiments

For all NLP experiments, our implementation is based on the GLUE (Wang et al., 2018a) experiments in the Transformers repository (Wolf et al., 2019) from Huggingface.[12] We used the Adam optimizer (Kingma & Ba, 2014) for every training set up, with a learning rate of 1e-5, and an epsilon value of 1e-8. We trained for 3 epochs for all *NLI tasks, maintaining a batch size of 92. All above hyperparameters were selected using a grid search; we kept other hyperparameters unaltered from the original HuggingFace repository. Each experiment was performed on a single Quadro RTX 8000 GPU.

A.7. Hyperparameters for ImageNet

We trained our ImageNet models using v3-512 TPU pods. For EfficientNet (Tan & Le, 2019), we used RandAugment data augmentation (Cubuk et al., 2019) with 2 layers, and a magnitude of 28, for all model sizes. We trained our models using a batch size of 4096, a learning rate of 0.128, and kept other hyperparameters the same as in (Tan & Le, 2019). We trained for 350 epochs for all dataset sizes - so when training on 20% or 40% of ImageNet (or a smaller dataset), we scaled the number of optimization steps accordingly. For ResNet (He et al., 2016), we used a learning rate of 0.1, a batch size of 8192, and trained for 90 epochs.

A.8. Qualitative Analysis of SNLI

Table 9 shows some examples removed and retained by AFLITE on the NLI dataset.

[12] https://github.com/huggingface/transformers

Table 9. Examples from SNLI, removed (top) and retained (bottom) by AFLite. As is evident, the retained instances are slightly more challenging and capture more nuanced semantics in contrast to the removed instances. Removed instances also exhibit larger word overlap, and many other artifacts found in Gururangan et al. (2018). Two examples per label are shown, the AFLite-filtered dataset contains many more `neutral` examples, as opposed to those labeled as `contradiction`.

REMOVED BY AFLITE		
Premise	**Hypothesis**	**Label**
A woman, in a green shirt, preparing to run on a treadmill.	A woman is preparing to sleep on a treadmill.	`contradiction`
The dog is catching a treat.	The cat is not catching a treat.	`contradiction`
Three young men are watching a tennis match on a large screen outdoors.	Three young men watching a tennis match on a screen outdoors, because their brother is playing.	`neutral`
A girl dressed in a pink shirt, jeans, and flip-flops sitting down playing with a lollipop machine.	A funny person in a shirt.	`neutral`
A man in a green apron smiles behind a food stand.	A man smiles.	`entailment`
A little girl with a hat sits between a woman's feet in the sand in front of a pair of colorful tents.	The girl is wearing a hat.	`entailment`

RETAINED BY AFLITE		
Premise	**Hypothesis**	**Label**
People are throwing tomatoes at each other.	The people are having a food fight.	`entailment`
A man poses for a photo in front of a Chinese building by jumping.	The man is prepared for his photo.	`entailment`
An older gentleman speaking at a podium.	A man giving a speech	`neutral`
A man poses for a photo in front of a Chinese building by jumping.	The man has experience in taking photos.	`neutral`
People are waiting in line by a food vendor.	People sit and wait for their orders at a nice sit down restaurant.	`contradiction`
Number 13 kicks a soccer ball towards the goal during children's soccer game.	A player passing the ball in a soccer game.	`contradiction`