

Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets

Yuxiang Wu^{†*} Matt Gardner^{‡*} Pontus Stenetorp[†] Pradeep Dasigi[§]
[†] University College London

yuxiang.wu, p.stenetorp@cs.ucl.ac.uk

[‡] Microsoft Semantic Machines [§] Allen Institute for AI

mattgardner@microsoft.com, pradeepd@allenai.org

Abstract

Natural language processing models often exploit spurious correlations between task-independent features and labels in datasets to perform well only within the distributions they are trained on, while not generalising to different task distributions. We propose to tackle this problem by generating a debiased version of a dataset, which can then be used to train a debiased, off-the-shelf model, by simply replacing its training data. Our approach consists of 1) a method for training *data generators* to generate high-quality, label-consistent data samples; and 2) a filtering mechanism for removing data points that contribute to spurious correlations, measured in terms of *z-statistics*. We generate debiased versions of the SNLI and MNLI datasets,¹ and we evaluate on a large suite of debiased, out-of-distribution, and adversarial test sets. Results show that models trained on our debiased datasets generalise better than those trained on the original datasets in all settings. On the majority of the datasets, our method outperforms or performs comparably to previous state-of-the-art debiasing strategies, and when combined with an orthogonal technique, product-of-experts, it improves further and outperforms previous best results of SNLI-hard and MNLI-hard.

1 Introduction

Natural Language Processing (NLP) datasets inevitably contain biases that are unrelated to the tasks they are supposed to represent. These biases are usually artifacts of the annotation processes, task framing, or design decisions (Schwartz et al., 2017; Geva et al., 2019; Liu et al., 2021). Such biases often manifest as spurious correlations between simple features of the data points and their

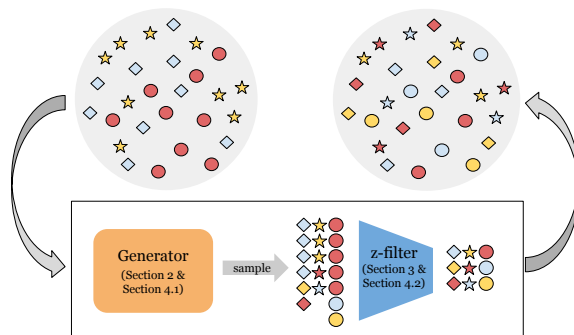


Figure 1: Overview of our dataset bias mitigation approach. We minimise spurious correlations between labels (represented by the shapes of data points) and task-independent features (represented by their colours) with our proposed data generation pipeline.

labels (Gardner et al., 2021). Trained models can exploit these spurious correlations to correctly predict the labels of the data points within the same distributions as those they are trained on, but fail to generalise to other distributions within the same tasks. Consequently, the models risk modelling the datasets, but not the tasks (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019; Schuster et al., 2019).

We address this issue by *adjusting* existing dataset distributions to mitigate the correlations between task-independent features and labels. First, we train *data generators* that generate high quality data samples in the distribution of existing datasets (Section 2). Then, we identify a set of simple features that are known to be task-independent, and use the theoretical framework (i.e., *z-statistics*) proposed by Gardner et al. (2021) to measure correlations between those features and the labels (Section 3.1). Finally, we adjust the distribution of the generated samples by post-hoc filtering (Section 3.2) to remove the data points that contribute to high *z-statistics* with task-independent features, or finetuning the data generator (Section 4.1) to make such data points less likely. Unlike prior *model-*

* Work done while at the Allen Institute for AI.

¹All our code and the generated datasets are available at <https://github.com/jimmycode/gen-debiased-nli>.

centric approaches to mitigate spurious correlations (Belinkov et al., 2019a,b; Clark et al., 2019; He et al., 2019; Karimi Mahabadi et al., 2020) that define new training objectives or model architectures, our approach has the advantage of keeping the objective and the model fixed, as we only alter the training data.

To evaluate our approach, we use the task of Natural Language Inference (NLI), which offers a wide range of datasets (including challenge datasets) for various domains. We generate debiased SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) distributions and evaluate the generalisability of models trained on them to out-of-distribution hard evaluation sets (Gururangan et al., 2018; McCoy et al., 2019), and the adversarial attack suite for NLI proposed by Liu et al. (2020b). Furthermore, we compare our method to strong debiasing strategies from the literature (Belinkov et al., 2019b; Stacey et al., 2020; Clark et al., 2019; Karimi Mahabadi et al., 2020; Utama et al., 2020; Sanh et al., 2021; Ghaddar et al., 2021).

Our results show that models trained on our debiased datasets generalise better than those trained on the original datasets to evaluation sets targeting hypothesis-only biases (by up to 2.8 percentage points) and syntactic biases (by up to 13.3pp), and to a suite of adversarial tests sets (by up to 4.2pp on average). Since our contributions are orthogonal to model-centric approaches, we show that when combined with product-of-experts (Karimi Mahabadi et al., 2020), our method yields further improvements and outperforms previous state-of-the-art results of SNLI-hard and MNLI-hard. Finally, we train stronger and larger pretrained language models with our debiased datasets, and demonstrate that the performance gain by our method generalises to these larger models.

2 Generating High-Quality Data Samples

First, we need to train a data generator G to generate data samples automatically. Our goal for the data generator is to model the true distribution as well as possible so that we can generate valid and high-quality data samples.

2.1 Finetuning Pretrained Language Model to Generate NLI Samples

We finetune a pretrained language model on the NLI datasets to serve as our data generator. We

choose GPT-2 because it is a powerful and widely-used autoregressive language model, and it can be easily adapted to generate the premise, label, and hypothesis of an instance sequentially.

Given an NLI dataset \mathcal{D}_0 , the training objective is to minimise the following negative log-likelihood loss of generating the premise-label-hypothesis sequence, in that order:

$$\begin{aligned} \mathcal{L}_{MLE} &= - \sum_{i=1}^{|\mathcal{D}_0|} \log p(P^{(i)}, l^{(i)}, H^{(i)}) \\ &= - \sum_{i=1}^{|\mathcal{D}_0|} \log p(P^{(i)})p(l^{(i)}|P^{(i)})p(H^{(i)}|l^{(i)}, P^{(i)}), \end{aligned} \quad (1)$$

where $P^{(i)}$, $l^{(i)}$ and $H^{(i)}$ are the premise, label and hypothesis respectively.²

2.2 Improving Data Generation Quality

We find that samples generated by a generator trained with only \mathcal{L}_{MLE} often contain ungrammatical text or incorrect label. In this section, we introduce two techniques to improve data quality.

2.2.1 Unlikelihood Training to Improve Label Consistency

We observe poor *label consistency* in samples generated by a generator trained with vanilla \mathcal{L}_{MLE} objective – given a generated sample $(\tilde{P}, \tilde{H}, \tilde{l})$, the label \tilde{l} often does not correctly describe the relationship between \tilde{P} and \tilde{H} . To alleviate this issue, we apply *unlikelihood training* (Welleck et al., 2020) to make generating such label inconsistent instances less likely.

First we perturb the label to construct negative samples (P, H, l') where $l' \neq l$ for each sample in the dataset. Then we apply a token-level unlikelihood objective on the hypothesis tokens:

$$\begin{aligned} \mathcal{L}_{\text{consistency}} &= \\ &= - \sum_{i=1}^{|\mathcal{D}_0|} \sum_{t=1}^{|H|^{(i)}} \log(1 - p(H_t^{(i)}|l'^{(i)}, P^{(i)}, H_{<t}^{(i)})). \end{aligned}$$

This objective decreases the probability of generating H when given an incorrect label l' , hence improves the label consistency at generation time.

²In our preliminary study, we found the factorization order premise-label-hypothesis in Eq. (1) performs better than hypothesis-label-premise and premise-hypothesis-label.

We combine \mathcal{L}_{MLE} and $\mathcal{L}_{\text{consistency}}$ to finetune our generator G with

$$\mathcal{L}_G = \mathcal{L}_{MLE} + \lambda \mathcal{L}_{\text{consistency}},$$

where λ is a hyperparameter that balances the two objectives. We can randomly sample from the trained generator to obtain a large amount of the synthetic data $\mathcal{D}_G \sim G$.

2.2.2 Filtering Based on Model Confidence

We add a consistency filtering step (Lewis et al., 2021; Bartolo et al., 2021) to further improve the quality of the generated dataset. We train an NLI model M with the original dataset \mathcal{D}_0 to filter out samples in which M has low confidence:

$$\hat{\mathcal{D}}_G = \{(P, H, l) \in \mathcal{D}_G \mid p_M(l|P, H) > \tau\},$$

where τ is a confidence threshold. We found that the filtered out data samples generally had ungrammatical text or incorrect labels.

3 Mitigating Spurious Correlations using *z*-filtering

We now define a method to reject samples that contribute to the high spurious correlations between task-independent features of the samples and their labels. Our approach is based on the theoretical framework proposed by Gardner et al. (2021) to measure these correlations, known as *z*-statistics. Our filtering method, called *z*-filtering (Section 3.2), will serve as the basis to construct debiased datasets in Section 4.

3.1 Identifying and Measuring Spurious Correlations

As a first step towards addressing spurious correlations, we need to be able to quantify them. We start by selecting a set of task-independent features – features that give away the labels and allow models to exploit them without actually solving the task. For NLI, we choose the following features: 1) unigrams and bigrams; 2) hypothesis length and hypothesis-premise length ratio; 3) lexical overlap between hypothesis and premise; 4) the predictions of a BERT-base (Devlin et al., 2019) hypothesis-only model.³ These features capture various biases identified in prior work, including contradiction word biases, lexical overlap bias (McCoy et al., 2019), and hypothesis-only bias (Gururangan et al.,

³See Appendix B for detailed descriptions of the features.

2018; Poliak et al., 2018). Note that our method does not rely on the specific choice of features, and one can easily add alternative features that should not be correlated with the labels.

Following Gardner et al. (2021), we assume there should be no correlation between each of these features and the class labels. More formally, for any feature x from our feature set \mathcal{X} , $p(l|x)$ should be uniform over the class labels l . We define $\hat{p}(l|x) = \frac{1}{n} \sum_{j=1}^n l^j$ to be the empirical expectation of $p(l|x)$ over n samples containing x . Then we compute the standardised version of *z*-statistics to quantify its deviation from the uniform distribution for each feature x and label l :

$$z^*(x, l) = \frac{\hat{p}(l|x) - p_0}{\sqrt{p_0(1 - p_0)/n}}, \quad (2)$$

where p_0 is the probability of uniform distribution ($p_0 = 1/3$ in NLI tasks with three labels).

These *z*-statistics scores can be used to identify the most biased features for each label l – we select k features with the highest *z*-statistic to define the *biased features* set $\mathcal{B}_{\mathcal{D}}(l)$. Table 12 shows examples of these biased features on SNLI.

3.2 *z*-filtering

To mitigate the biases in the dataset, we propose *z*-filtering, an algorithm that iteratively selects and filters instances from a dataset \mathcal{D}' to build a debiased dataset \mathcal{Z} . At each step, we find the set of biased features $\mathcal{B}_{\mathcal{Z}}(l)$ on the partially constructed \mathcal{Z} . We then select a new batch of samples from \mathcal{D}' and filter out the samples that contain these biased features. This process is applied iteratively until it has exhausted all samples from \mathcal{D}' . It removes the samples that contribute to the spurious correlations in \mathcal{D}' , thus it finds a debiased subset $\mathcal{Z}(\mathcal{D}') \subset \mathcal{D}'$. We denote the removed samples as $\mathcal{Z}^-(\mathcal{D}')$. The full *z*-filtering algorithm is illustrated in Algorithm 1.

Optionally, one can initialise \mathcal{Z} with a seed dataset $\mathcal{D}_{\text{seed}}$. In this case, the samples from \mathcal{D}' are only added to \mathcal{Z} when they do not contain the biased features of $\mathcal{D}_{\text{seed}}$. Thus it can be seen as a data-augmentation technique targeted to debias a given dataset. We refer to it as *conditional z*-filtering and denote the produced debiased dataset as $\mathcal{Z}(\mathcal{D}'|\mathcal{D}_{\text{seed}})$.

Algorithm 1: z-filtering algorithm.

Data: input dataset \mathcal{D}' [with optional seed dataset \mathcal{D}_{seed}]
Result: debiased dataset \mathcal{Z} and the rejected samples \mathcal{Z}^-
 $\mathcal{Z} \leftarrow \emptyset$ (or $\mathcal{Z} \leftarrow \mathcal{D}_{seed}$);
 $\mathcal{Z}^- \leftarrow \emptyset$;
for sample batch $\mathcal{D}'_t \subset \mathcal{D}'$ **do**
 compute or update z-statistics
 $z^*(x, l|\mathcal{Z}), \forall x \in \mathcal{X}$ of \mathcal{Z} ;
 find the biased features $\mathcal{B}_{\mathcal{Z}}(l), \forall l \in \{\text{entailment, neutral, contradiction}\}$;
 foreach instance $I = (P, H, l) \in \mathcal{D}'_t$ **do**
 get the features f of the instance I ;
 if $f \cap \mathcal{B}_{\mathcal{Z}}(l) = \emptyset$ **then**
 $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{I\}$;
 else
 $\mathcal{Z}^- \leftarrow \mathcal{Z}^- \cup \{I\}$;
 end
 end
end

4 Constructing Debiased NLI Datasets via Data Generation

We use z-filtering in two ways: 1) to further finetune G (the one trained in Section 2.2.1 with consistency unlikelihood) with an objective that downweights samples that should be rejected (Section 4.1); 2) to post-hoc filter the generated samples to obtain debiased datasets (Section 4.2).

4.1 Learning to Generate Unbiased Samples

The generator G can learn to exploit task-independent features during its finetuning stage (Section 2), causing the synthetic data $\hat{\mathcal{D}}_G$ to contain many spurious correlations. While it is tempting to apply z-filtering to remove these spurious correlations from $\hat{\mathcal{D}}_G$, we find that this will lead to the removal of majority of the generated data. For example, when the generator is finetuned on SNLI, z-filtering removes around 85% of $\hat{\mathcal{D}}_{G_{SNLI}}$.⁴ This leads to a very inefficient data generation process to mitigate the spurious correlations.

To alleviate this issue, we can incorporate the debiasing objectives into the training of the generator, so that the samples produced by the generator

⁴This is also strong confirmation that these biases are problematic, as the generative model easily finds them and relies on them during data generation. Conducting naive data augmentation with $\hat{\mathcal{D}}_{G_{SNLI}}$ will strengthen the spurious correlations.

are more likely to be accepted by the z-filtering process. More specifically, we can encourage the model to generate $\mathcal{Z}(\mathcal{D}_0)$, while discouraging it from generating $\mathcal{Z}^-(\mathcal{D}_0)$. For the latter part, we again apply an unlikelihood training objective \mathcal{L}_{UL} to *unlearn* $\mathcal{Z}^-(\mathcal{D}_0)$. Hence, the overall debiasing training objective is:

$$\mathcal{L}_{debias} = \mathcal{L}_{MLE}(\mathcal{Z}(\mathcal{D}_0)) + \alpha \mathcal{L}_{UL}(\mathcal{Z}^-(\mathcal{D}_0))$$

where α is a hyperparameter.

A naive use of an unlikelihood objective on all tokens gives the model mixed signals for good tokens and leads to ungrammatical, degenerate outputs. To avoid this degeneracy, we apply the unlikelihood loss only to tokens that contribute to biased features. Concretely, for each token I_t^- of instance $I^- \in \mathcal{Z}^-(\mathcal{D}_0)$, we define a mask m_t as

$$m_t = \begin{cases} 0, & \text{if } I_t^- \text{ contributes to } \mathcal{B}_{\mathcal{Z}}(l_{I^-}) \\ 1, & \text{otherwise.} \end{cases}$$

where $\mathcal{B}_{\mathcal{Z}}(l_{I^-})$ represent the biased features corresponding the label of I^- .

For biases towards unigram and bigram features (as defined in Section 3.1), we consider only the corresponding tokens to be relevant (i.e., $m_t = 0$ if I_t^- is part of the unigram or the bigram). For biases towards other features (e.g. length of the hypothesis), we consider all the tokens on the hypothesis to be relevant. The unlikelihood training objective is defined as follows:

$$\begin{aligned} \mathcal{L}_{UL}(\mathcal{Z}^-(\mathcal{D}_0)) &= \sum_{I' \in \mathcal{Z}^-(\mathcal{D}_0)} \mathcal{L}_{UL}(I'), \\ \mathcal{L}_{UL}(I') &= - \sum_{t=1}^{|I'|} \log(m_t p(I'_t | I'_{<t})) \\ &\quad + (1 - m_t)(1 - p(I'_t | I'_{<t})). \end{aligned}$$

We further finetune G with \mathcal{L}_{debias} to obtain a new generator G^* , that is trained to generate more unbiased data samples. We then randomly sample from G^* and conduct data filtering (Section 2.2.2) to obtain a large set of high-quality debiased data samples $\hat{\mathcal{D}}_{G^*}$.

4.2 Combining with z-filtering to Construct the Debiased NLI Datasets

Given the original dataset \mathcal{D}_0 and the synthetic dataset $\hat{\mathcal{D}}_{G^*}$, our goal is produce a large-scale unbiased dataset \mathcal{D}^* . There are various ways to do

this given that we can either apply conditional z-filtering, or simply z-filter both \mathcal{D}_0 and $\hat{\mathcal{D}}_{G^*}$ and merge them. We explore the following options:

1. **Z-Augmentation (Z-Aug)** $\mathcal{Z}(\hat{\mathcal{D}}_{G^*}|\mathcal{D}_0)$: we keep the original dataset as is, and augment it by conducting conditional z-filtering on $\hat{\mathcal{D}}_{G^*}$ using \mathcal{D}_0 as seed dataset.
2. **Parallel z-filter (Par-Z)** $\mathcal{Z}(\mathcal{D}_0) \cup \mathcal{Z}(\hat{\mathcal{D}}_{G^*})$: we conduct z-filtering on \mathcal{D}_0 and $\hat{\mathcal{D}}_{G^*}$ separately, and then merge them.
3. **Sequential z-filter (Seq-Z)** $\mathcal{Z}(\hat{\mathcal{D}}_{G^*}|\mathcal{Z}(\mathcal{D}_0))$: we first conduct z-filtering on \mathcal{D}_0 , then conduct conditional z-filtering on $\hat{\mathcal{D}}_{G^*}$ with $\mathcal{Z}(\mathcal{D}_0)$ as seed dataset.

5 Experiments

5.1 Experimental Setup

Source Datasets We select the two most widely used NLI datasets SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) as our original datasets. Prior work (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019) found various annotation artifacts in them, hence they serve as good use cases for constructing debiased datasets.

Evaluation Datasets For the hypothesis-only bias, we use the challenge sets SNLI-hard (Gururangan et al., 2018) and MNLI-hard (Williams et al., 2018), which were produced by filtering the test set with a hypothesis-only model (Section 5.2). For syntactic biases, we follow previous work and use HANS (McCoy et al., 2019) for evaluation (Section 5.3). In addition, we evaluate on the adversarial test benchmark introduced by Liu et al. (2020b) (Section 5.4). This benchmark covers a wide range of adversarial attacks, which will give a more complete picture of what spurious correlations the debiasing methods tackle.

Generating Debiased Datasets We conduct debiased data generation for SNLI and MNLI *separately*. For SNLI, we use the proposed method described in Section 4.1 to train a generator G_{SNLI}^* . Then we randomly sample a large number of instances from the generator to construct $\mathcal{D}_{G_{\text{SNLI}}^*}$. The samples are filtered with a strong NLI model M trained on SNLI to obtain $\hat{\mathcal{D}}_{G_{\text{SNLI}}^*}$. Finally, different options (Section 4.2) can be adopted to merge the synthetic data with the original data $\mathcal{D}_{\text{SNLI}}$ to construct debiased versions of SNLI. The same

Options	$\mathcal{D}_0 = \mathcal{D}_{\text{SNLI}}$	$\mathcal{D}_0 = \mathcal{D}_{\text{MNLI}}$
Original \mathcal{D}_0	549,367	382,702
Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_0)$	1,142,475	744,326
Par-Z $\mathcal{Z}(\mathcal{D}_0) \cup \mathcal{Z}(\hat{\mathcal{D}}_{G^*})$	933,085	740,811
Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_0))$	927,906	744,200

Table 1: Data size of the constructed debiased datasets for SNLI and MNLI.

procedure is used to produce debiased datasets for MNLI, by simply replacing the original dataset with MNLI. We choose GPT-2 large and Roberta-large as the pretrained language models for G^* and M respectively.⁵ The size of the constructed debiased datasets are listed in Table 1.

NLI Model Training Since our method directly debiases the training data itself, we keep the model and training objective fixed and only replace the training data with our generated debiased datasets. For comparability with previous work (Karimi Mahabadi et al., 2020; Utama et al., 2020; Sanh et al., 2021), we train BERT-base (Devlin et al., 2019) on our debiased datasets. The NLI models are trained with ordinary *cross-entropy* classification loss, and the training hyperparameters are listed in Appendix A. We run our experiments five times and report the average and standard deviation of the scores.⁶ We also conduct statistical significance testing using a 2-tailed t-test at 95% confidence level.

State-of-the-art Debiasing Models We compare our method with the following three state-of-the-art debiasing models on each of our evaluation datasets. **Product-of-Experts** (He et al., 2019; Karimi Mahabadi et al., 2020) ensembles a bias-only model’s prediction b_i with the main model’s p_i using $p'_i = \text{softmax}(\log p_i + \log b_i)$. This ensembling enforces that the main model focuses on the samples that the bias-only model does not predict well. **Learned-Mixin** (Clark et al., 2019) is a variant of PoE that introduces a learnable weight for the bias-only model’s prediction. **Regularized-conf** (Utama et al., 2020) uses confidence regularisation to retain the in-distribution performance while conducting model debiasing.

⁵On one A100 GPU, training the generator takes around 24 hours and generating the samples takes roughly 35 hours for each dataset.

⁶With the exception of our PoE experiments which single run, as hyperparameter tuning for PoE is costlier.

Method (model w/ data)	SNLI	SNLI-hard
Prior debiasing strategies trained on SNLI		
AdvCls (Belinkov et al., 2019a)*	83.56	66.27
Ens. AdvCls (Stacey et al., 2020)*	84.09	67.42
DFL (Karimi Mahabadi et al., 2020)*	89.57	83.01
PoE (Karimi Mahabadi et al., 2020)*	90.11	82.15
<hr/>		
BERT-base w/ $\mathcal{D}_{\text{SNLI}}$ baseline	90.45	80.34 \pm 0.46
Models trained on our debiased datasets		
BERT-base w/ Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{SNLI}})$	90.67	<u>81.78</u> \pm 0.53
BERT-base w/ Par-Z $\mathcal{Z}(\mathcal{D}_{\text{SNLI}} \cup \mathcal{Z}(\hat{\mathcal{D}}_{G^*}))$	88.11	<u>82.81</u> \pm 0.37
BERT-base w/ Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{SNLI}}))$	88.08	<u>82.82</u> \pm 0.15
<hr/>		
Combining PoE with our debiased datasets		
BERT-base + PoE w/ $\mathcal{D}_{\text{SNLI}}$	90.25	82.92
BERT-base + PoE w/ Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{SNLI}}))$	87.65	84.48

Table 2: Accuracy on SNLI and SNLI-hard. * are reported results and underscore indicates statistical significance against the baseline. Training on our debiased SNLI datasets significantly boosts the performance on SNLI-hard compared to the baseline, and it improves further when combined with PoE.

Combining PoE with Our Debiased Datasets

Our approach changes the training data distribution instead of the model’s training objective, and hence is orthogonal to prior work method-wise. We also report the results of combining PoE with our proposed method, simply by training a PoE model on our debiased datasets. We adapt the PoE implementation by Karimi Mahabadi et al. (2020), and we follow their approach to conduct hyperparameter tuning for PoE.⁷ The hyperparameters of the PoE models are reported in Table 10 of Appendix A.

5.2 Hypothesis-only Bias in NLI

Gururangan et al. (2018) found that, on SNLI and MNLI, a model that only has access to the hypothesis can perform surprisingly well, which indicates that the datasets contain hypothesis-only bias. To alleviate this problem, SNLI-hard and MNLI-hard (Gururangan et al., 2018) subsets were constructed by filtering the test set with a hypothesis-only model and only accepting those that the hypothesis-only model predicts incorrectly. We examine whether our method successfully mitigates the hypothesis-only bias in NLI, by evaluating the models trained with our debiased datasets on SNLI-hard and MNLI-hard.

Results on SNLI-hard Table 2 shows the results of our method on SNLI and SNLI-hard. The results show that, compared to training on SNLI, training with our debiased datasets significantly improves the performance on SNLI-hard. The

debiased dataset produced by Seq-Z achieves a 2.48% gain in accuracy on SNLI-hard compared to the SNLI baseline, whereas Z-Aug improves both SNLI and SNLI-hard accuracy.

Results on MNLI-hard Table 3 shows the results of our method on MNLI-matched (MNLI-m) and MNLI-mismatched (MNLI-mm), and their corresponding hard sets. We use the development sets of MNLI-hard reconstructed by (Karimi Mahabadi et al., 2020) to develop our methods. To comply with the submission limit of MNLI leaderboard system, we select the best checkpoint among the five runs using the development set, and report its test set performance in Table 3.

The results show that BERT-base models trained on our debiased MNLI datasets outperform the models trained on the original MNLI by a large margin on the MNLI-hard sets. In particular, the Z-Aug version of the debiased datasets gives a 2.72% and 2.76% gain in accuracy on MNLI-m hard and MNLI-mm hard respectively, and outperforms the previous state-of-the-art on MNLI-m, MNLI-mm, and MNLI-mm hard.

Combining PoE with Our Debiased Datasets

We investigate the combination of our method and PoE, to see if the two orthogonal techniques can work together to achieve better performance. Since hyperparameter tuning of PoE is costly, we choose the best version of the debiased dataset (Seq-Z for SNLI and Z-Aug for MNLI) using the development set accuracy, and train PoE with it. The results are listed in the last rows of Table 2 and Table 3. We can find that, on both SNLI and MNLI, combining PoE with our debiased dataset yields further improvements on SNLI-hard, MNLI-m hard, and MNLI-mm hard, outperforming previous state-of-the-art results on all three datasets.

5.3 Syntactic Bias in NLI

McCoy et al. (2019) show that NLI models trained on MNLI can exploit syntactic heuristics present in the data, such as lexical overlap, subsequence, and constituent features. They introduce HANS, an evaluation dataset that contains examples where the syntactic heuristics fail. To test whether our method mitigates the syntactic biases in NLI, we evaluate models trained on our debiased datasets on HANS. If our debiased dataset contains less syntactic bias than the original dataset, the model would not exploit the syntactic heuristics and thus perform better on HANS. Due to the high variance

⁷<https://github.com/rabeehk/robust-nli>

Method (model w/ data)	MNLI-m		MNLI-mm		MNLI-m hard		MNLI-mm hard	
	dev	test	dev	test	dev	test	dev	test
Prior debiasing strategies trained on MNLI								
PoE (Karimi Mahabadi et al., 2020)*	84.58	84.11	84.85	83.47	78.02	76.81	79.23	76.83
Learned-Mixin (Clark et al., 2019)*	80.5	79.5	81.2	80.4	-	79.2	-	78.2
Regularized-conf (Utama et al., 2020)*	84.6	84.1	85.0	84.2	-	78.3	-	77.3
BERT-base Main PoE+CE (Sanh et al., 2021)*	83.32	-	83.54	-	-	77.63	-	76.39
BERT-base w/ $\mathcal{D}_{\text{MNLI}}$ baseline	83.87	84.11	84.22	83.51	76.39 ± 0.64	75.88	77.75 ± 0.45	75.75
Models trained on our debiased datasets								
BERT-base w/ Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{MNLI}})$	84.72	85.12	85.14	84.09	78.95± 0.76	78.60	80.29± 0.54	78.51
BERT-base w/ Par-Z $\mathcal{Z}(\mathcal{D}_{\text{MNLI}}) \cup \mathcal{Z}(\hat{\mathcal{D}}_{G^*})$	82.48	83.27	82.95	82.95	78.88 ± 0.80	79.19	80.02 ± 0.62	78.49
BERT-base w/ Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{MNLI}}))$	82.55	83.41	82.70	83.17	78.88 ± 0.83	79.19	79.65 ± 0.44	78.44
Combining PoE with our debiased dataset								
BERT-base + PoE w/ $\mathcal{D}_{\text{MNLI}}$	84.39	84.69	84.25	83.75	78.37	77.54	79.45	78.33
BERT-base + PoE w/ Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{MNLI}})$	85.22	85.38	85.72	84.53	80.49	80.03	81.52	79.28

Table 3: Accuracy on MNLI-matched (MNLI-m), MNLI-mismatched (MNLI-mm), MNLI-matched hard, and MNLI-mismatched hard. * are reported results and underscore indicates statistical significance against the baseline. Training on our debiased MNLI datasets significantly boosts the performance on MNLI-matched hard and MNLI-mismatched hard. When combined with PoE, our method improves further and outperforms previous methods.

Method	HANS
Methods trained on SNLI	
BERT-base Attention (Stacey et al., 2021)*	58.42
Roberta-large w/ AFLite (Bras et al., 2020)*	59.6
Roberta-base w/ TAILOR (Ross et al., 2021)*	70.5
Methods trained on MNLI	
Learned-Mixin (Clark et al., 2019)*	64.00
Learned-Mixin+H (Clark et al., 2019)*	66.15
PoE (Karimi Mahabadi et al., 2020)*	66.31 ± 0.6
DFL (Karimi Mahabadi et al., 2020)*	69.26 ± 0.2
PoE+CE (Sanh et al., 2021)*	67.9
Regularized-conf (Utama et al., 2020)*	69.1 ± 1.2
E2E Self-debias (Ghaddar et al., 2021)*	71.2± 0.2
Models trained on our debiased datasets	
Roberta-base w/ $\mathcal{D}_{\text{SNLI}}$	65.32 ± 2.22
Roberta-base w/ Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{SNLI}}))$	66.87± 1.47
BERT-base w/ $\mathcal{D}_{\text{MNLI}}$ baseline	54.36 ± 2.56
BERT-base w/ Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{MNLI}})$	62.57± 5.91
BERT-base w/ Par-Z $\mathcal{Z}(\mathcal{D}_{\text{MNLI}}) \cup \mathcal{Z}(\hat{\mathcal{D}}_{G^*})$	65.11± 5.62
BERT-base w/ Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{MNLI}}))$	67.69± 3.53
BERT-base + PoE w/ $\mathcal{D}_{\text{MNLI}}$ (baseline)	63.40
BERT-base + PoE w/ Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{MNLI}})$	68.75
Roberta-large w/ $\mathcal{D}_{\text{MNLI}}$	75.74 ± 2.82
Roberta-large w/ Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{MNLI}})$	78.65± 2.26

Table 4: Results on HANS (McCoy et al., 2019). * are reported results and underscore indicates statistical significance against the baseline. BERT-base trained on our debiased MNLI datasets performs significantly better than the one trained on the original MNLI, and it improves further when combined with PoE. Roberta-large also benefits from training on our debiased dataset.

of the scores on HANS, we run five times for each experiment (except PoE), and report the average and standard deviation of the scores.

Results on HANS Table 4 shows the results on HANS. The results are categorised into three sections according to the training data: SNLI, MNLI, and our debiased datasets. The results of models trained on our debiased MNLI datasets show strong improvements: compared to the original MNLI, our debiased MNLI datasets obtain up to a 13.33% gain in HANS accuracy. Our Seq-Z variant achieves 67.69% accuracy, which is comparable with strong PoE baseline (Karimi Mahabadi et al., 2020; Sanh et al., 2021). Our method also further improves PoE models: the BERT-base PoE model trained on our Z-Aug MNLI outperforms the one trained on MNLI by 5.3%. Additionally, training Roberta-large (Liu et al., 2019) on our debiased dataset introduces 2.9 points accuracy gain on HANS, indicating that the performance gain by our debiased dataset can generalise to larger and stronger models (more on this in Section 5.5).

5.4 Adversarial Tests for Combating Distinct Biases in NLI

Liu et al. (2020b) find that debiasing methods often tie to one particular known bias and it is non-trivial to mitigate multiple NLI biases at the same time. They introduce a suite of test datasets for NLI models that targets various aspects of robustness, including partial input heuristics (PI), logical infer-

	PI-CD	PI-SP	IS-SD	IS-CS	LI-LI	LI-TS	ST	Avg.
Data-augmentation heuristics proposed by Liu et al. (2020b)								
Text Swap*	71.7	72.8	63.5	67.4	86.3	86.8	66.5	73.6
Sub (synonym)*	69.8	72.0	62.4	65.8	85.2	82.8	64.3	71.8
Sub (MLM)*	71.0	72.8	64.4	65.9	85.6	83.3	64.9	72.6
Paraphrase*	72.1	74.6	66.5	66.4	85.7	83.1	64.8	73.3
BERT-base w/ $\mathcal{D}_{\text{MNLI}}$ baseline	70.3 \pm 0.5	73.7 \pm 1.4	53.5 \pm 2.3	64.8 \pm 1.4	85.5 \pm 0.9	81.6 \pm 1.4	69.2 \pm 0.8	71.2 \pm 0.8
Models trained on our debiased datasets								
BERT-base w/ Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{MNLI}})$	73.1 \pm 0.9	76.1 \pm 1.2	61.8 \pm 6.1	69.1 \pm 1.3	86.9 \pm 0.6	83.1 \pm 0.9	70.1 \pm 0.5	74.3 \pm 1.3
BERT-base w/ Par-Z $\mathcal{Z}(\mathcal{D}_{\text{MNLI}}) \cup \mathcal{Z}(\hat{\mathcal{D}}_{G^*})$	72.0 \pm 0.9	78.7 \pm 1.2	64.5 \pm 5.8	70.7 \pm 1.7	88.5 \pm 0.7	82.6 \pm 0.3	69.6 \pm 1.0	75.2 \pm 1.4
BERT-base w/ Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{MNLI}}))$	71.7 \pm 0.9	77.8 \pm 1.2	66.9 \pm 3.7	71.1 \pm 0.7	89.1 \pm 1.0	82.3 \pm 0.9	69.3 \pm 0.8	75.4 \pm 0.8

Table 5: Results on the NLI adversarial test benchmark (Liu et al., 2020b). We compare with the data augmentation techniques investigated by Liu et al. (2020b). * are reported results and underscore indicates statistical significance against the baseline. Training on our debiased MNLI datasets significantly improves the performance on majority of the categories (PI-CD, PI-SP, IS-SD, IS-CS, LI-LI) and on average.

ence ability (LI), and stress test (ST).⁸ Several data augmentation strategies were investigated by Liu et al. (2020b): 1) text swap: swapping the premise and hypothesis in the original data; 2) word substitution: replacing words in the hypothesis with synonyms or generations from a masked language model; 3) paraphrase: using back translation to paraphrase the hypothesis.

We compare our approach with their data-augmentation heuristics, and the results are shown in Table 5. Comparing with the MNLI baseline, our debiased MNLI datasets lead to better performance across all categories, which indicates that our method successfully mitigates various distinct biases simultaneously. All three variants of our debiased datasets outperform the data augmentation heuristics by Liu et al. (2021), which demonstrates the efficacy of our method when compared against manually designed heuristics.

5.5 Generalisation to Larger Pretrained Language Models

Since our method mitigates the spurious correlations in the dataset, not the model, our approach is model-agnostic and has the potential to benefit larger future models. To test this hypothesis, we train stronger and more modern models than BERT with our debiased datasets, and see if it can still improve the performance. More specifically, we choose Roberta-base, Roberta-large (Liu et al., 2019), and Albert-xxlarge (Lan et al., 2020), train them with Seq-Z SNLI and Z-Aug MNLI.

The results in Table 6 show that: 1) these larger models achieve better generalisation performance than BERT-base, which agrees with Bhargava et al.

(2021); Bowman (2021); 2) training on our debiased datasets can still improve the performance of these models, yielding an average 2.30%, 1.23%, 1.13% gain for Roberta-base, Roberta-large and Albert-xxlarge respectively. This indicates that our method generalises to larger pretrained language models and could potentially enhance future models.

	Test data	Original	Debiased	Δ
Roberta-base	SNLI-hard	82.02 \pm 0.24	83.71 \pm 0.31	1.69
	MNLI-m hard	81.74 \pm 0.44	83.14 \pm 0.25	1.40
	MNLI-mm hard	81.93 \pm 0.30	83.12 \pm 0.24	1.19
	HANS	71.17 \pm 2.95	76.15 \pm 1.52	4.98
	Adv.Test avg	77.63 \pm 0.49	79.89 \pm 0.38	2.26
Roberta-large	SNLI-hard	83.61 \pm 0.31	85.09 \pm 0.32	1.48
	MNLI-m hard	85.44 \pm 0.62	85.69 \pm 0.24	0.25
	MNLI-mm hard	85.37 \pm 0.63	85.94 \pm 0.21	0.57
	HANS	75.74 \pm 2.82	78.65 \pm 2.26	2.91
	Adv.Test avg	80.92 \pm 0.46	81.86 \pm 0.31	0.94
Albert-xxlarge	SNLI-hard	83.59	84.82	1.23
	MNLI-m hard	86.42	86.40	-0.02
	MNLI-mm hard	86.38	86.82	0.44
	HANS	76.32	79.05	2.73
	Adv.Test avg	81.91	83.18	1.27

Table 6: Performance gain when training larger models with our debiased datasets. Underscore indicates statistical significance against the baseline that is trained on the original datasets. For evaluation on SNLI-hard, the models are trained with SNLI or our debiased Seq-Z SNLI; for other evaluation datasets, the models are trained with MNLI or our debiased Z-Aug MNLI. Albert-xxlarge is experimented with one run due to its higher training cost.

⁸Details of the subcategories are described in Appendix C.

6 Related Work

Spurious Correlations in Datasets The issue of spurious correlations in datasets between labels and simple input features has recently received significant attention (Gururangan et al., 2018; Poliak et al., 2018; Belinkov et al., 2019a; Karimi Mahabadi et al., 2020). It has been shown that this issue is often inherent in the data annotation process, caused by biases in the framing of the task (Schwartz et al., 2017), noisy annotations (Chen et al., 2016), or personal (Geva et al., 2019) or group-level (Liu et al., 2021) annotator biases. Gardner et al. (2021) provide a theoretical framework for analyzing spurious correlations, which we use to define our filtering mechanism in Section 3.2.

Debiasing NLI Models Much prior work follows a *model-centric* approach towards mitigating biases in NLI models – they propose novel model architectures or training objectives to ensure that the models do not exploit the shortcuts presented by the dataset biases. At the representation level, Belinkov et al. (2019a,b) introduce an adversarial architecture to debias hypothesis representations to tackle hypothesis-only bias (Gururangan et al., 2018), and Stacey et al. (2020) strengthen the debiasing by using multiple adversarial classifiers. Zhou and Bansal (2020) use HEX projection to project the representation to the space orthogonal to the biased features to debias the model. At the model level, Clark et al. (2019); He et al. (2019); Karimi Mahabadi et al. (2020) propose methods based on Product-of-Expert (PoE) (Hinton, 2002) for mitigating biases by ensembling a biased-only model with a main model. Utama et al. (2020) propose the use of confidence regularization to improve out-of-distribution performance while retaining in-distribution accuracy.

Debiasing NLI Datasets Ross et al. (2021) introduce TAILOR, a semantically-controlled perturbation method for data augmentation based on a small number of manually defined perturbation strategies. Bras et al. (2020) propose AFLite, a dataset filtering method that learns feature representations with a model and conduct adversarial filtering based on model predictions. Unlike these approaches, our method requires no manually-written perturbation heuristics and is model-agnostic, hence it is more generally applicable.

Generative Data Augmentation Several works investigate generative data augmentation techniques to improve model robustness in other areas. Yang et al. (2020) conduct generative data augmentation for commonsense reasoning and show that it can improve out-of-domain generalisation. Lee et al. (2021) trains a generator to generate new claims and evidence for debiasing fact verification datasets like FEVER (Thorne et al., 2018). Schick and Schütze (2021) exploit large pretrained language models to generate semantic textual similarity datasets. Bartolo et al. (2021) improve robustness of question answering models by generating adversarial dataset.

7 Conclusions

To address the issue of spurious correlations between task-independent features and labels in NLI datasets, we propose methods to generate label-consistent data and then filter out instances from existing datasets that contribute to those spurious correlations; thereby generating debiased datasets. Models trained on our debiased versions of the SNLI and MNLI datasets generalise better than the equivalent model trained on the original datasets to a large suite of test sets focusing on various kinds of known biases. Future work in this direction includes investigating whether our techniques are applicable to tasks beyond NLI.

Acknowledgments

The authors would like to thank Max Bartolo, Alexis Ross, Doug Downey, Jesse Dodge, Pasquale Minervini, and Sebastian Riedel for their helpful discussion and feedback.

References

- Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. [Improving question answering model robustness with synthetic adversarial data generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019a. [Don’t take the premise for granted: Mitigating artifacts in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

- Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019b. [On adversarial removal of hypothesis-only bias in natural language inference](#). In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota. Association for Computational Linguistics.
- Prajwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. [Generalization in NLI: Ways \(not\) to go beyond simple heuristics](#). In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Samuel R. Bowman. 2021. [When combating hype, proceed with caution](#). *CoRR*, abs/2110.08300.
- 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.
- Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. [Adversarial filters of dataset biases](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.
- Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. [A thorough examination of the CNN/Daily Mail reading comprehension task](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.
- 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.
- 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.
- 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.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. [Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets](#). 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.
- Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. [End-to-end self-debiasing framework for robust NLU training](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929, Online. Association for Computational Linguistics.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that require simple lexical inferences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- He He, Sheng Zha, and Haohan Wang. 2019. [Unlearn dataset bias in natural language inference by fitting the residual](#). In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.
- Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.
- 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.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [ALBERT: A lite BERT for self-supervised](#)

- learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3181–3185.
- Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.
- Haochen Liu, Joseph Thekinen, Sinem Mollaoglu, Da Tang, Ji Yang, Youlong Cheng, Hui Liu, and Jiliang Tang. 2021. Toward annotator group bias in crowdsourcing. *ArXiv preprint*, abs/2110.08038.
- Tianyu Liu, Zheng Xin, Baobao Chang, and Zhifang Sui. 2020a. HypoNLI: Exploring the artificial patterns of hypothesis-only bias in natural language inference. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6852–6860, Marseille, France. European Language Resources Association.
- Tianyu Liu, Zheng Xin, Xiaoan Ding, Baobao Chang, and Zhifang Sui. 2020b. An empirical study on model-agnostic debiasing strategies for robust natural language inference. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 596–608, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- 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.
- Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, Brussels, Belgium. 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.
- Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of NLI models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA*, pages 6867–6874. AAAI Press.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *ArXiv preprint*, abs/2107.07150.
- Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. Learning from others’ mistakes: Avoiding dataset biases without modeling them. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics.
- Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.
- Joe Stacey, Yonatan Belinkov, and Marek Rei. 2021. Supervising model attention with human explanations for robust natural language inference. *ArXiv preprint*, abs/2104.08142.
- Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8281–8291, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. [Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Haohan Wang, Da Sun, and Eric P. Xing. 2019. [What if we simply swap the two text fragments? A straightforward yet effective way to test the robustness of methods to confounding signals in nature language inference tasks](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA*, pages 7136–7143. AAAI Press.

Sean Welleck, Ilya Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. [Neural text generation with unlikelihood training](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. [G-daug: Generative data augmentation for commonsense reasoning](#). In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 1008–1025. Association for Computational Linguistics.

Xiang Zhou and Mohit Bansal. 2020. [Towards robustifying NLI models against lexical dataset biases](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Online. Association for Computational Linguistics.

A Hyperparameters

A.1 Hyperparameters of Our Proposed Method

Hyperparameter	Value
learning rate	1e-5
batch size	24
epoch	5
optimiser	Adam
Adam ϵ	1e-6
Adam (β_1, β_2)	(0.9, 0.999)
learning rate scheduler	constant
max sequence length	128
pretrained model	GPT-2 large
device	Nvidia A100
λ	0.5
α	1.0

Table 7: Hyperparameters for training the generator G^* .

Hyperparameter	Value
number of samples from G_{SNLI}^*	5,000,000
number of samples from G_{MNLI}^*	4,000,000
data filtering threshold τ	0.95
data filtering model	Roberta-large
z-filtering number of biased features	20

Table 8: Hyperparameters of the data generation pipeline.

Hyperparameter	Value
learning rate	1e-5
batch size	32
epoch	5
optimiser	Adam
Adam ϵ	1e-6
Adam (β_1, β_2)	(0.9, 0.999)
learning rate scheduler	constant with warmup
warm up steps	2000
max sequence length	128
pretrained model	BERT-base
device	Nvidia A100
early stop patience	3 epochs

Table 9: Hyperparameters for training the NLI models.

A.2 Hyperparameter Tuning of PoE

The learning objective of PoE is defined as follows:

$$\mathcal{L}_{\text{PoE}} = \sum_{i=1}^{|\mathcal{D}|} CE(l_i, p'_i) + \gamma CE(l_i, b_i),$$

where CE stands for cross-entropy loss, l_i is the label, and γ is a hyperparameter. $p'_i = \text{softmax}(\log p_i + \beta \log b_i)$ is the ensemble of the main model’s prediction p_i , and the bias-only model’s prediction b_i weighted by a hyperparameter β .

We find that the result of PoE is very sensitive to the hyperparameters β and γ . Following Karimi Mahabadi et al. (2020), we conduct grid search for the two hyperparameters, with $\beta \in \{0.05, 0.1, 0.2, 0.4, 0.8, 1.0, 2.0\}$ and $\gamma \in \{0.05, 0.1, 0.2, 0.4, 0.8, 1.0\}$. The best hyperparameters found for each evaluation dataset is listed in Table 10.

Train data	Eval. data	β	γ
SNLI	SNLI-hard	2.0	0.4
Seq-Z SNLI	SNLI-hard	2.0	0.4
MNLI	MNLI-m hard	0.8	1.0
	MNLI-mm hard	2.0	0.4
	HANS	2.0	0.8
Z-Aug MNLI	MNLI-m hard	2.0	0.4
	MNLI-mm hard	2.0	0.8
	HANS	2.0	1.0

Table 10: Best hyperparameters found for PoE models with different training and evaluation datasets.

B Task-independent Features

We list the chosen set of task-independent features that we aim to mitigate in this work in Table 11. Note that our method does not depend on the choice of task-independent features. One can easily add their own features in the future to mitigate newly-identified spurious correlations.

Table 12 shows the most salient task-independent features (ranked by z-statistics) in SNLI and our debiased SNLI dataset. It shows that the correlation between task-independent features and labels is massively reduced, dropping from over 400 to roughly 17. These results verify that our method successfully mitigates the spurious correlations in the dataset.

Feature	Description
Unigrams & Bigrams	All unigrams and bigrams. The n-grams from premise and hypothesis are treated separately.
Hypothesis length	Number of tokens in the hypothesis.
Hypothesis-premise length ratio	Number of tokens in hypothesis divided by number of tokens in the premise.
Lexical overlap	Ratio of tokens in the hypothesis that overlap with the premise.
Hypothesis-only model’s prediction	We train a hypothesis-only model on the original dataset and use its prediction as a feature.
Null feature	A dummy feature added for <i>all</i> instances to avoid skewed label distribution.

Table 11: Descriptions of the features used to debias the datasets in Section 3.

SNLI		Debiased SNLI (Seq-Z)	
Biased feature	z-statistics	Biased feature	z-statistics
Entailment			
hypo-only-pred=0	422.1	theres@hypothesis	17.5
lex-overlap > 0.8	123.3	hypo-len < 5	17.4
full-lex-overlap	117.3	full-lex-overlap	17.4
outside@hypothesis	102.2	politician@hypothesis	17.4
lex-overlap > 0.9	90.4	speaking@hypothesis	17.4
Neutral			
hypo-only-pred=1	436.1	championship@hypothesis	15.3
for a@hypothesis	63.6	living room@hypothesis	15.2
his@hypothesis	56.8	many men@hypothesis	15.2
friends@hypothesis	55.6	green suit@hypothesis	15.2
tall@hypothesis	52.7	are wearing@hypothesis	15.2
Contradiction			
hypo-only-pred=2	433.9	nothing@hypothesis	17.0
sleeping@hypothesis	92.9	hypo-only-pred=2	16.9
is sleeping@hypothesis	68.7	at home@hypothesis	16.9
nobody@hypothesis	68.4	is no@hypothesis	16.9
no@hypothesis	62.7	york yankees@hypothesis	16.9

Table 12: Top-5 biased features with the highest z-statistics on SNLI (left) and debiased Seq-Z SNLI (right) for each label class.

C Description of Adversarial Test (Liu et al., 2020b) Subcategories

The adversarial test benchmark (Liu et al., 2020b) includes the following subcategories from various sources:

- PI-CD: classifier detected partial-input (Gururangan et al., 2018).
- PI-SP: HypoNLI (Liu et al., 2020a) dataset that tackles surface patterns heuristics.
- IS-SD: syntactic diagnostic dataset HANS (McCoy et al., 2019).
- IS-CS: lexically misleading instances constructed by Nie et al. (2019).
- LI-LI: lexical inference test by (Naik et al., 2018; Glockner et al., 2018).
- LI-TS: text-fragment swap test by swapping the premise and hypothesis (Wang et al., 2019; Minervini and Riedel, 2018).

- ST: an aggregation of word-overlap (ST-WO), negation (ST-NE), length mismatch (ST-LM), and spelling errors (ST-SE) tests in (Naik et al., 2018).

D Visualisation of z-statistics

Following Gardner et al. (2021), we visualise the statistics of the features on both SNLI and our debiased SNLI (Seq-Z) dataset in Fig. 2.⁹ Comparing the two plots, it confirms that our method successfully suppresses the spurious correlations in the dataset.

E Ablation Study

Data	Size	SNLI	SNLI-hard
Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{SNLI}}))$	928k	88.08	82.82 \pm 0.15
Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{Z}(\mathcal{D}_{\text{SNLI}}))$	549k	87.59	82.35 \pm 0.46
Seq-Z $\mathcal{Z}(\hat{\mathcal{D}}_G \mathcal{Z}(\mathcal{D}_{\text{SNLI}}))$	549k	88.15	82.20 \pm 0.23
$\mathcal{D}_{\text{SNLI}} \cup \hat{\mathcal{D}}_{G^*}$	2577k	90.85	81.99 \pm 0.47
$\mathcal{D}_{\text{SNLI}} \cup \hat{\mathcal{D}}_G$	3717k	90.83	80.82 \pm 0.27
Z-Aug $\mathcal{Z}(\hat{\mathcal{D}}_{G^*} \mathcal{D}_{\text{SNLI}})$	1142k	90.67	81.78 \pm 0.53
$\mathcal{D}_{\text{SNLI}} \cup \hat{\mathcal{D}}_{G^*}$	1142k	90.72	81.45 \pm 0.52
$\mathcal{D}_{\text{SNLI}} \cup \hat{\mathcal{D}}_G$	1142k	90.67	80.85 \pm 0.27
$\mathcal{Z}(\hat{\mathcal{D}}_{G^*})$	808k	88.44	81.28 \pm 0.57
$\mathcal{Z}(\hat{\mathcal{D}}_{G^*})$	549k	88.12	80.67 \pm 0.41
$\hat{\mathcal{D}}_{G^*}$ (w/ filter)	549k	88.59	80.41 \pm 0.50
\mathcal{D}_{G^*} (wo/ filter)	808k	75.65	76.67 \pm 0.83
\mathcal{D}_{G^*} (wo/ filter)	549k	75.43	76.05 \pm 0.49
$\mathcal{Z}(\mathcal{D}_{\text{SNLI}})$	127k	84.93	80.52 \pm 1.03
original SNLI $\mathcal{D}_{\text{SNLI}}$	549k	90.45	80.34 \pm 0.46

Table 13: Ablation study conducted on SNLI and SNLI-hard.

⁹We sample 10% of the points under the $z = 10.0$ curve to compress the figure, but it may still be slow to render the figures because the number of points is still large.

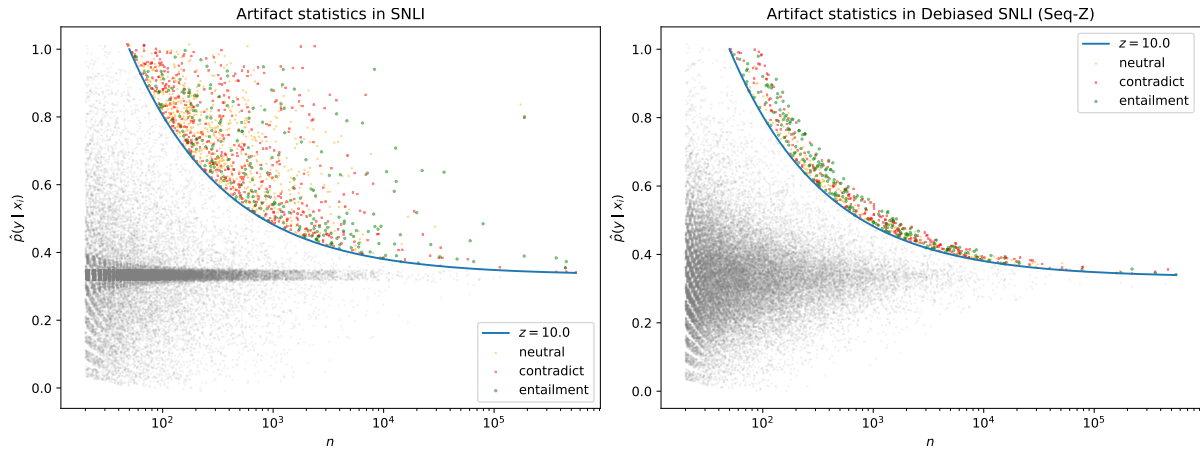


Figure 2: Statistics of the features on SNLI and our debaised SNLI (Seq-Z).

F Generated Samples of Debaised Dataset

Table 14 and Table 15 show generated samples in the debaised SNLI and MNLI datasets respectively. The samples are quite diverse and the quality is reasonably good, which demonstrates the effectiveness of our quality ensuring techniques presented in Section 2.2.

Premise	Hypothesis	Label
Thanksgiving dinner is a fun time for everyone.	The dinner is a fun event.	entailment
A father is letting his toddler drink from his glass.	A toddler is having a drink	entailment
Hair stylist performing a haircut.	A hailer is performing surgery	contradiction
Then there are two men in white shirts, one of which is holding a cigarette and the other an open book.	Two men sit at a conference table with a book and a cigarette.	neutral
Three men playing basketball on a court with an audience in the background.	Three people playing basketball	entailment
Six children, boys and girls, jumping into a swimming pool.	Six children are jumping into a pool	entailment
Three girls jump for joy in front of a building.	The kids are sitting on their front steps.	contradiction
The child in the green one piece suit is running in the playground.	The child is playing outside	entailment
View of an intersection with city buses and a police car.	The intersection is surrounded by vehicles.	entailment
The man on the yellow basketball team tries to score while the men on the opposing team try to block his shot.	Two men on different teams are competing in a game of a male is trying to score while other men on the opposing one defend his basket in basketball	entailment
Young child wearing orange shirt eating a ice cream cone.	A child eats ice cream at the ice cream stand.	neutral
Five people standing in front of a shopping center.	Five people outside the building	entailment
He's taking a break after a long workout.	He is taking a break from his workout	entailment
Two men are sitting on a couch, playing music together.	The two people play guitars.	neutral
Everyone is out enjoying the winter weather and having fun with their children.	Everyone is out enjoying the summer	contradiction
Many people walking through a city street.	There are a group of people in Times Square.	neutral
A woman in a black dress walks down the street.	a person in dresses walks	entailment
Four children, riding unicycles, are on a sidewalk in front of a brick building.	Four children ride unicycles on the sidewalk	entailment
Four kids playing soccer in a field.	The children played with bubbles.	contradiction
MADISON, Wis. (AP) — The man in the white jersey and orange visor threw the ball for the two boys in uniforms with blue jerseys.	A man in white is throwing a ball to two boys in blue uniforms.	entailment
Mikhail Kasyapkin, who plays Bart on The Simpsons, is talking to a woman.	The woman tells him to stop making couples sit	neutral
Shutterstock photo of a woman with a heart tattoo on her calf.	A woman with a pumpkin tattoo on her back	contradiction
Three women and a man sing their hearts out in the microphone.	A group singing	entailment
With so many people on the beach, the woman in yellow has to make a quick decision.	Many people are at a beach, one has to make a decision	entailment
Celebrants are walking with American flags.	People are walking.	entailment
Customer examines flowers at a market.	A customer examines flowers.	entailment
He is in the air on his skateboard.	A guy is in a tree.	contradiction
thousands of people enjoying a fireworks show.	There is an audience for a show.	entailment
Bicyclists in a race, with a blue bike leaving the ground in the lead.	Bikers resting after a long ride.	contradiction
He has a pet bird in a cage, and it is sleeping.	He is walking the dogs.	contradiction

Table 14: Generated samples in the debiased SNLI datasets.

Premise	Hypothesis	Label
As I noted earlier, the board and the auditors should have a strategic alignment of interests.	The board should align to increase efficiency.	neutral
This story was originally published in Slate. For more on the U.S. role in that war, subscribe to Slate's Subscribe now!	The U.S. played very little part in the war.	neutral
Via Newsday's a poll finds that 84 percent of Americans think Monica Lewinsky should tell the truth about her encounter with Clinton.	A majority of the public thinks Lewinsky should come forward.	entailment
Violence among theatrical people, on the other hand, can be entertainingly savage, cf. All About Eve (1884) and The Mousetrap (1928).	There aren, always hasn't usually been oancy situation with violence among theatrical people because they don't have to work because it isn't employment.	contradiction
Nowhere in the book does Hatfield warn the reader that he has altered details or created composite characters to protect his sources.	Hatfield didn't inform the readers in any part in the book that the details of the altered information was to protect his sources	entailment
The young inhabitants are brought up knowing nothing else.	The young inhabitants have been brought up knowing of nothing.	entailment
The 5th floor of the Royal Palace is open to the public, with restricted access for foreign guests.	Foreign guest have restricted access in the royal palace for visitors.	entailment
Pulitzer Prizes are given to books, magazines, paintings, and sculpture. I admit I didn't have much reason to think that.	You won a prize when you eat blueberries at dinner. After all, most of the people don;t think that way.	contradiction neutral
In the past, Medicare's fiscal health has generally been gauged by the solvency of the HI trust fund projected over a 75-year period.	Medicare's soliesic fitness is displayed in the form of the surplus projected over a 50 year term.	contradiction
A case study where the only people interviewed were senior officials would be seen as a not-good case study, in contrast to one where the views of individuals at all levels affected was obtained.	If senior editors were interviewed they would not be considered the best examples for case studies.	entailment
If you've ever spent an evening plunging your wrists into ice water, you are an easy mark for devices that promise to relieve carpal tunnel syndrome.	People are easy marks for devices that may cure cat paral tunnel syndrome	entailment
It's a sign of a permanently altered world that natural blondness should have such sacred power no longer.	The people still believe blondness has a special significance.	contradiction
The Three-Arched Bridge, by Ismail Kadare, translated by John Hodgson (Arcade).	Ismail Marare translated The Three-Aral.	contradiction
Many of these organizations found themselves in an environment similar to the one confronting federal managers today-one in which they were called upon to improve performance while simultaneously reducing costs.	This was the only option for all their group.	neutral
The long-sought, the elusive, the elusive Jane Finn!	She is easily obtainable.	contradiction
And now, to-day, he puts forward a suggestion that he himself must have known was ridiculous.	He is making the ridiculous suggestion that himself must have been aware of.	entailment
Jupiter's moon, Callisto, has a thick atmosphere and is a good destination for a quiet tour.	Callisto's atmosphere makes for a pleasant journey to explore.	entailment
Founded in 1995, the Agora formed to address the enormous security challenges brought about by new computer, network, and Internet technologies.	The Agora was formed to address the challenge of nuclear proliferation.	contradiction
Just last week in The New Yorker, Malcolm Gladwell argued that Gen.	Just last week in Newsweek, Johnny Chung argued that Gen.	contradiction
Muller and most of the boys can be counted on not to cause any more than the normal pay-night disturbances.	Muller will not start a fist fight.	neutral
Don't call me Shirley.	My last name is Shirley and that is how I want to be referred to.	contradiction
The vast majority of the approximately 1,700 lawyers at LSC-funded programs around the country volunteer for only a single case, whether it is a class action suit, a simple civil rights case or a case involving a dangerous person.	There's no reason to get one or do the work otherwise.	neutral
The Promise Keepers talk far less about abortion and homosexuality than their critics and the media do.	They're surrounded far less with the issues that the media and other critics deal with.	entailment
It was Susan in his head.	Susan was telling him exactly to his surprise.	neutral
In 1782, after only a few years, the city decided to impose planning guidelines.	It took a few decades for 17 year-olds.	contradiction

Table 15: Generated samples in the debiased MNLi datasets.