HuaSLIM: Human Attention Motivated Shortcut Learning Identification and Mitigation for Large Language Models

Yuqi Ren and Deyi Xiong *
College of Intelligence and Computing, Tianjin University, Tianjin, China
{ryq20, dyxiong}@tju.edu.cn

Abstract

Large language models have made remarkable progress on a variety of NLP tasks. However, it has been found that they tend to rely on shortcut features that spuriously correlate with labels for prediction, which weakens their generalization on out-of-distribution samples. In this paper, we propose a human attention guided approach to identifying and mitigating shortcut learning, which encourages the LLM-based target model to learn relevant features. We define an attention-based measurement to capture both model and data bias and identify shortcut tokens by exploring both human and neural attention. In a self-distillation framework, we mitigate shortcut learning by dynamically adjusting the distillation temperature according to the detected shortcut tokens and estimated shortcut degree. Additionally, we utilize human attention as a supervisory signal to constrain large language models to pay more attention to relevant tokens. Experimental results on multiple NLP tasks show that our proposed method can effectively identify shortcut tokens, and significantly improve the robustness of large language models on OOD samples, while not undermining the performance on IID data.

1 Introduction

Large language models (LLMs), e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020), have achieved state-of-the-art performance in a wide variety of NLP tasks. However, recent studies show that these models often exploit spurious correlations between shortcut tokens and labels, rather than capture underlying semantics related to the target task (Utama et al., 2020b; McCoy et al., 2019; Gururangan et al., 2018). Such LLMs may suffer from the robustness issue when confronted with out-of-distribution (OOD) samples as spurious correlations (i.e., shortcut features) learned from the training data are usually absent in OOD samples (Wang and Culotta, 2021).

The main causes of shortcut learning include data bias usually caused by data crowdsourcing (Gururangan et al., 2018) and model bias towards learning simple features (Shah et al., 2020). Previous works measure the degree of shortcut learning often by data statistics and model interpretability methods (McCoy et al., 2019; Du et al., 2021). Particularly, they estimate the shortcut degree of each sample based on the tokens that are correlated with labels. However, they do not distinguish shortcut tokens (spuriously correlated tokens) from genuine correlated tokens (Winship and Morgan, 1999).

With identified shortcut tokens, various approached have been proposed to suppress LLM-based task-specific models learning shortcut features by adjusting the loss function to mitigate shortcut learning, such as sample re-weighting (Schuster et al., 2019), product of experts (Sanh et al., 2021). These methods have achieved significant improvements on OOD samples, but at the cost of undermining the performance of LLM-based task-specific models on independent and identically distributed (IID) samples (Utama et al., 2020a). Additionally, recent studies have found that these methods actually encode more biases to the inner representations of LLMs (Mendelson and Belinkov, 2021).

In this paper, we propose to address these issues via human attention that implicates the cognitive processing behaviour of human brains. With the aid of human attention, we want to encourage LLM-based task-specific models to learn relevant features, so as to improve the performance on both IID and OOD samples. Incorporating human attention into neural models can be regarded as a human-in-the-loop learning, where human feedback has proven capable of not only effectively improving both the accuracy and robustness of models, but also building strong interpretability and credibil-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2023, pages 12350–12365
July 9-14, 2023

ity for models (Wang et al., 2021; Stiennon et al., 2020). Additionally, human attention based approaches have been successfully applied in a range of NLP tasks recently, such as paraphrase generation (Sood et al., 2020), entity linking (Klie et al., 2020).

Encouraged by these, we propose **HuaSLIM**, a **Hu**man **a**ttention motivated **S**hortcut **L**earning **I**dentification and **M**itigation framework for large language models. Specifically, to identify shortcut tokens, we introduce an attention-based local mutual information metric that takes into account both lexical bias and model behavior bias to detect tokens highly correlated with certain labels. We then automatically distinguish spurious correlations from genuine correlations based on the orthogonal information between human attention-based correlation and neural attention-based correlation, instead of directly using the tokens that are highly correlated with labels. Intuitively, 'spurious' tokens are paid more attention to during model training, while less attention in human reading. For shortcut learning mitigation, we based HuaSLIM on self-distillation (Furlanello et al., 2018). We utilize the estimated shortcut learning degree of each sample to dynamically adjust the temperature in distillation, in the goal of softening the output distribution of the teacher model, thereby discouraging the reliance of LLM-based task-specific models on shortcut learning.

Additionally, we force LLM-based task-specific models to learn how humans understand language by simulating human reading behavior. Specifically, we introduce a new training objective that drives neural attention to fit human attention distribution. In this way, LLM-based task-specific models are trained to explicitly pay more attention to relevant tokens identified by human attention. To avoid the effect of attention heads playing different roles in Transformer (Clark et al., 2019b), we add an additional soft attention layer after the last layer of LLMs.

Human attention signals used in this paper are deduced from human gaze duration generated by EZ-Reader that has been widely used in the study of human reading process (Reichle et al., 2009, 2013). This avoid expensive collection of eye movement signals.

In a nutshell, our contributions are listed as follows:

- We introduce a shortcut learning degree measurement based on human attention to automatically identify shortcut tokens. Our analyses show that it can effectively distinguish spurious correlations from genuine correlations.
- We use the shortcut learning degree of samples to control the temperature in the self-distillation of LLMs, significantly improving their robustness.
- We propose a human attention guided shortcut learning mitigation method, which forces LLMs to shift attention from shortcut features to genuine features implicated by human attention.
- We conduct experiments on three NLP tasks: NLI, fact verification and paraphrase identification. Results suggest that proposed method can significantly improve the performance on both IID and OOD samples.

2 Related Work

Identification of Shortcut Learning. As shortcut learning has significantly hurt the robustness of neural models, a large number of studies have been dedicated to identifying the shortcut learning problem and understanding how neural networks exploit spurious correlations (Sagawa et al., 2020; Wang and Culotta, 2020; Wang et al., 2022). In early works, adversarial datasets are built to evaluate the generalization ability of neural models on OOD samples, such as HANS that evaluates whether NLI models adopt fallible syntactic heuristics (McCoy et al., 2019), Symmetrics that evaluates the effect of shortcut tokens on fact verification (Schuster et al., 2019). Data analysis and model interpretability analysis are also used to detect shortcut tokens that are considered highly correlated with final predictions by neural models, e.g., integrated gradient (Du et al., 2021), neural attention (Wang et al., 2022). Such extracted shortcut tokens facilitate the alleviation of the shortcut learning issue in neural models.

Mitigation of Shortcut Learning. A wide variety of model-centric approaches have been recently proposed to mitigate shortcut learning, e.g., explanation regularization (Liu and Avci, 2019), product of experts (Sanh et al., 2021), sample re-weighting (Schuster et al., 2019; Liu et al., 2021), confidence

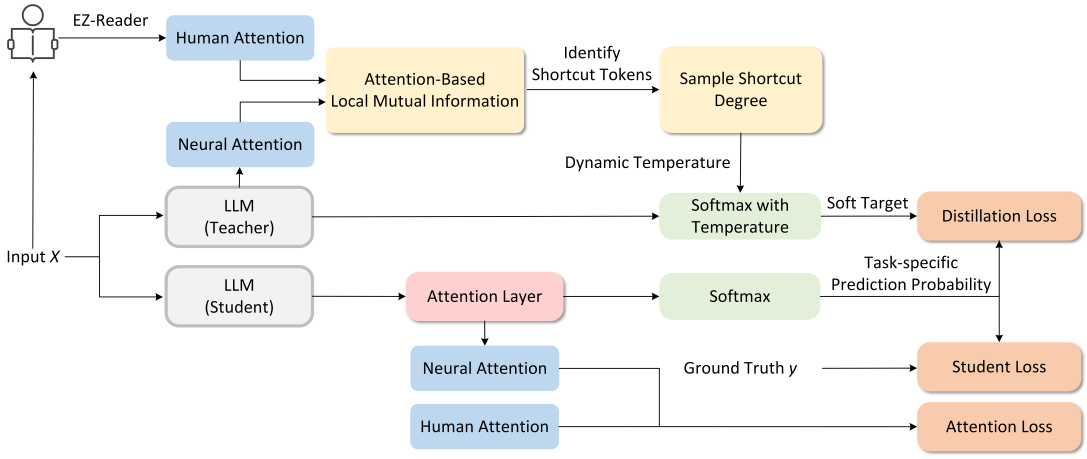

Figure 1: The diagram of the proposed HuaSLIM. It is based on self-distillation where the teacher and student model have the same model architecture. The teacher model is trained to measure the shortcut learning degree of samples with human attention. The estimated shortcut learning degree is integrated into the temperature coefficient to dynamically smooth the output distribution. The student model is trained to learn human attention to guide the model to pay attention to relevant features. An additional attention layer is employed to better model neural attention.

regularization (Utama et al., 2020a), etc. Data augmentation methods that aim at improving robustness have also been explored (Wu et al., 2022; Si et al., 2021). While most methods significantly improve the performance on OOD samples by mitigating shortcut learning, they may undermine the performance of IID samples (Mendelson and Belinkov, 2021). In addition to this, it is difficult to analyze whether LLM-based task-specific models obtain more robust features. Significantly different from previous shortcut learning mitigation approaches, we leverage human attention to learn robust and interpretable features and attempt to boost performance on both OOD and IID samples.

Human Attention in NLP. Human attention, tracked by human gaze signals and implicating the cognitive language comprehension process of human brains (Henderson, 2003; Rayner, 1978), has been attracting research interest in cognitive science. Integrating human attention into neural network models has been applied to a large number of natural language processing tasks, such as prediction of multiword expressions (Rohanian et al., 2017), paraphrase generation (Sood et al., 2020), machine reading (Li et al., 2018). In most works, human gaze signals are used as additional input features to enhance the performance of neural network models for NLP tasks (Klerke and Plank, 2019; Zhang and Zhang, 2019). Other studies regularize neural networks in a multi-task learning framework, where human attention prediction is treated as an auxiliary task (Barrett et al., 2018; Klerke et al., 2016). Unlike them, we use human attention as supervisory signals to constrain the neural attention of LLMs.

3 Methodology

Our HuaSLIM aims to use human attention to guide model training, introducing human prior knowledge and reasoning ability into LLM-based task-specific models, thereby improving the robustness of LLM-based task-specific models. The architecture of HuaSLIM is illustrated in Figure 1. We use human attention produced by EZ-reader (Reichle et al., 2003) to identify shortcut tokens. To quantitatively detect shortcut learning, we propose a human attention-based sample shortcut degree measurement. With estimated shortcut learning degree scores, we inhibit LLM-based task-specific models from making overconfident predictions for samples containing shortcut features by dynamically adjusting temperature. To further force LLM-based task-specific models to focus on relevant features, we minimize the distance between human attention and neural attention with an attention loss.

3.1 Human Attention

In cognitive science, human gaze duration is usually used to track human attention to tokens during reading process (Lindsay, 2020). However, building a real human eye-tracking dataset is very expensive. Instead, we use the cognitively inspired

model EZ-reader (Reichle et al., 2003), which has proven an effective way to closely resemble real eye movement signals (Eberle et al., 2022), to simulate human attention for different NLP tasks. To match LLMs, we feed tokenized inputs into EZ-reader. Token-level $gaze\ durations$ generated by EZ-reader are hence considered as human attention in this work.

3.2 Identifying Shortcut Learning

In general, a shortcut token co-occurs more frequently with a target label than other tokens in training data (Gururangan et al., 2018) and neural models tend to learn simple features like this (Shah et al., 2020). Most shortcut learning identification methods capture the tokens that are highly correlated with labels by analyzing data distribution or model behavior, then identify the top-K most important tokens as shortcut tokens. In this paper, we propose an attention-based Local Mutual Information (LMI) (Evert, 2005) metric to identify shortcut tokens. LMI is usually used to measure the correlation between a token and a particular label in data statistics (Schuster et al., 2019; Du et al., 2021). The proposed attention-based metric can take into account both lexical bias and model behavior bias to capture the token-label correlation as we replace the token frequency term in traditional LMI with attention weights. Specifically, the co-occurrence number $\mathrm{count}(t, y)$ of token t with label y in traditional LMI is replaced by the sum of attention weights $\mathrm{attention}(t, y)$ between token t and label y in the training data. The proposed attention-based metric ALMI between token t and label y, is calculated as follows:

$$\mathrm{ALMI}(t, y) = p(t, y) \cdot \log(\frac{p(y|t)}{p(y)}) \quad (1)$$

where $p(t, y) = \frac{\mathrm{attention}(t,y)}{|D|}$, $p(y|t) = \frac{\mathrm{attention}(t,y)}{\mathrm{attention}(t)}$, $p(y) = \frac{\mathrm{attention}(y)}{|D|}$. $\mathrm{attention}(t, y)$ is the sum of attention weights between token t and label y. $\mathrm{attention}(t)$ is the sum of attention weights of '[CLS]' token in the last layer of LLM for token t. $\mathrm{attention}(y)$ is the sum of attention weights for all tokens in the samples labeled y. $|D|$ is the sum of attention weights for all tokens in the training data.

Obviously, the correlations detected in the above way contain both spurious and genuine correlations, since they are both strongly associated with labels. The genuine correlations have a causal effect on model predictions, while the spurious correlations cannot causally affect model predictions although they are highly correlated with specific labels (Wang et al., 2022).

We hence need to recognize the spurious correlations from the obtained correlations. Intuitively, humans rarely rely on shortcut words for comprehension and reasoning, focusing instead on relevant words. Inspired by this, we propose to identify genuine correlations according to human attention, and detect shortcut tokens according to the difference between neural attention based correlations and human attention based correlations. Particularly, we obtain a correlation list based on human attention and a correlation list based on neural attention on the same data via the proposed attention-based LMI. Then, we use MinMax to normalize the correlation scores from the two lists to the range of [0,1]:

$$I_{\mathrm{scale}} = \frac{I - \min(I)}{\max(I) - \min(I)} \quad (2)$$

where I is the correlation scores based on human attention or neural attention. In this way, we obtain normalized correlation scores for both neural attention and human attention: I^n_{scale}, I^h_{scale}. We then calculate $(I^n_{\mathrm{scale}} - I^h_{\mathrm{scale}})/I^n_{\mathrm{scale}}$ as token-level shortcut degree and re-rank tokens according to their degree scores. Intuitively, tokens with higher shortcut degree scores indicate that they are treated more important in model prediction but less important in human reading. Therefore, they are more likely to be shortcut tokens.

With estimated token-level shortcut degree, we further propose a measurement to calculate sample-level shortcut degree. Specifically, we consider top-N tokens in terms of their token-level shortcut degree as shortcut tokens, and normalize their shortcut degree scores to the range of [0,1]. Given a training sample x_i, the sum of token-level shortcut degree scores in the sample is defined as the sample shortcut degree β_i. In the following subsections, we utilize β_i to guide the model distillation.

3.3 Self-Distillation for Mitigating Shortcut Learning

Our shortcut learning mitigation is based on self-distillation (Furlanello et al., 2018), where the teacher model and the student model have identical architecture. In traditional knowledge distillation (Hinton et al., 2015), temperature T of soft target is used to control the softening degree of the

output probability of the teacher model. A higher temperature makes the distribution smoother, thus increasing the difficulty of model training (Li et al., 2022).

For training samples with a high shortcut degree, we increase the temperature to soften the target distribution, improving the learning difficulty of the student model on them, so as to inhibit LLMs to make overconfident predictions. Based on the teacher model and shortcut degree of each sample, we smooth the soft target by dynamically adjusting the temperature coefficient:

$$s_{i,j} = \frac{\exp(P^t_{i,j}/(T+\beta_i))}{\sum_{l=1}^{L}\exp(P^t_{i,l}/(T+\beta_i))} \quad (3)$$

where L denotes the number of labels, and P^t is the output of the teacher model. The temperature coefficient corresponding to sample x_i is the sum of the constant temperature T and sample shortcut degree β_i, which is dynamically adjusted with the sample shortcut degree.

3.4 Attention Layer

Since attention heads in transformer encode different semantic information (Clark et al., 2019b; Vig and Belinkov, 2019), it is difficult to determine which heads should be supervised by human attention will benefit the most. We stack an additional attention layer that is the same as soft-attention (Shen and Lee, 2016) which explicitly generates token-level attention weights, over the last layer of the transformer. The calculation of the stacked attention $\boldsymbol{\alpha}^n$ is as follows:

$$\boldsymbol{\alpha}^n = \text{softmax}(v^T \tanh(W_{\text{att}}\mathbf{H}^s + b_{\text{att}})) \quad (4)$$

where W_{att}, b_{att}, v are trainable parameters, and $\mathbf{H}^s$ denotes the hidden state of last layer in student model. $\boldsymbol{\alpha}^n$ indicates the degree of the importance of each token to model prediction after softmax normalization. The final sentence representation of the student model can be formulated as:

$$\mathbf{h}^s = \sum_{i=1}^{N} \alpha_i^n \mathbf{H}_i^s \quad (5)$$

We then obtain the normalized prediction probability P^s of the student model by softmax function:

$$P^s = \text{softmax}(W_s\mathbf{h}^s + b_s) \quad (6)$$

3.5 Training Objective

The training objective of the self-distillation in HuaSLIM is similar to that of traditional knowledge distillation framework, including the distillation loss for learning the scaled output of the teacher model and the student loss for learning the ground truth. The role of distillation loss is to transfer knowledge from teacher model to student model. The total loss of self-distillation is computed as follows:

$$\mathcal{L}_{\text{dis}} = -\sum_{k=1}^{K}((1-\lambda)y_k \log P_k^s + \lambda s_k \log P_k^s) \quad (7)$$

where K denotes the number of samples and y denotes the ground-truth label. Hyperparameter λ denotes the balancing weight for controlling the importance of each training objective.

To encourage the student model to focus on more relevant tokens, we use human attention as the inductive bias of neural attention. Therefore, we introduce an additional loss to fit human attention, allowing the student model to learn prior knowledge from humans. The additional training objective is to minimize the mean square error between neural attention of the stacked additional attention layer and human attention:

$$\mathcal{L}_{\text{att}} = \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}(\alpha_i^n - \alpha_i^h)^2 \quad (8)$$

where a_i^h denotes the human attention score for token i in sentence k and N is the number of tokens in the sample.

The final training objective that we minimize during training is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{att}} \quad (9)$$

During training, only the parameters of the student model are updated, while the parameters of the teacher model are fixed. For inference, we therefore only use the student model.

4 Experiments

4.1 Datasets

We used three datasets to evaluate the proposed HuaSLIM.

MNLI The first dataset we used is MNLI (Williams et al., 2018), a natural language inference dataset for training models that classify the entailment of a given pair of premise and hypothesis into classes (e.g., entailment, neutral, and contradiction). MNLI consists of development sets and test sets in two different domains, one is MNLI-m, which matches the domain of the training set, while the other is MNLI-mm, which is out of the domain of the training set. Additionally, the adversarial dataset HANS (McCoy et al., 2019) was used to test the robustness of LLMs on OOD samples. HANS is constructed based on the strong correlations between lexical overlap and entailment labels, which is widely used for robustness evaluation (Utama et al., 2020a; Du et al., 2021).

FEVER The second dataset is FEVER (Thorne et al., 2018), a dataset for the fact verification task predicting whether the claim sentences are labeled in the context of evidence sentences as support, refutes, or not enough information. There are two adversarial datasets associated with FEVER: Symmetric v1 and v2 (Sym1 and Sym 2), which test the model's reliance on claim-only bias (e.g., negative tokens such as 'not' are associated with the refutes label) that performs above the 'majority' baseline (Schuster et al., 2019). All claim-evidence pairs in these datasets are created manually, and shortcut features are distributed across labels.

Quora Question Pairs We chosen Quora Question Pairs (QQP) as the third dataset. It is a dataset for paraphrase identification task predicting whether pairs of questions are semantically duplicate or non-duplicate. We used a QQP subset PAWS (Paraphrase Adversaries from Word Scrambling) that consists of question pairs that have lexical overlap biases (Zhang et al., 2019) to evaluate the OOD performance of models. Most samples in this dataset are labeled as non-duplicate. Since neural models usually heavily rely on lexical overlap features, their performance on this dataset is worse than the random baseline (Zhang et al., 2019). We evaluated the performance of LLM-based task-specific models on the duplicate and non-duplicate samples separately, following Utama et al. (2020b).

4.2 Large Language Models

We conducted experiments on two LLMs to examine the effectiveness of HuaSLIM: BERT-base (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), both of which are from Hugging Face Transformers.[1] We followed the standard setting of sentence pair classification tasks, in which two sentences are connected into one input by '[SEP]' token. As mentioned in Section 3.4, we stacked an additional attention layer over the top layer of the two LLMs. We hence utilized the output of added attention layer for prediction, instead of using the hidden state of special token '[CLS]'.

4.3 Baselines

Sample Re-weighting Its main idea is assigning higher weights to hard samples, making LLM-based task-specific model pay more attention to difficult features, so as to improve the robustness of the model (Schuster et al., 2019; Utama et al., 2020b). In the first step, a bias-only model that trained by the hand-crafted features that based on the task-specific knowledge is trained to measure how well the sample prediction given only the biased features. In the second step, probability p_b obtained by bias-only model is used to indicate the shortcut degree of the sample. Then, adjusting the loss function with the shortcut degree to reduce the contribution of shortcut sample to LLM-based task-specific model:

$$\mathcal{L} = -(1 - p_b)y \cdot \log p_d \tag{10}$$

where p_d is the prediction probability of LLM-based task-specific model.

Product of Experts The purpose of product of experts is to integrate the bias-only model to train a debiased model (He et al., 2019; Clark et al., 2019a). First, a bias-only model is trained to capture biases in the training data. We then optimize the loss of this method, which is the combination of losses of both the debiased and bias-only model. The ensemble loss of product of experts as follows:

$$\mathcal{L} = -p_d \cdot \log \texttt{softmax}(\log p_b + \log p_d) \tag{11}$$

Product of experts prevent LLM-based task-specific model from learning shortcut features by reducing the gradient of shortcut samples in the training data, while it also compromises the model's ability to learn from these samples.

Confidence Regularization This method encourages LLM-based task-specific model to give lower confidence for shortcut samples by regularizing the

[1] https://huggingface.co/transformers/pretrained_models.html

Methods	MNLI			FEVER			QQP			
	dev-m$^{\diamondsuit}$	dev-mm$^{\diamondsuit}$	HANS	dev$^{\diamondsuit}$	Sym1	Sym2	dev$^{\diamondsuit}_{\text{dup}}$	dev$^{\diamondsuit}_{\neg\text{dup}}$	PAWS$_{\text{dup}}$	PAWS$_{\neg\text{dup}}$
BERT-base	84.2	83.4	61.5	85.2	55.3	63.1	88.3	91.5	85.2	23.2
with Sample Re-weighting	83.5†	81.3	69.2†	84.3‡	56.4‡	64.9‡	85.5	91.9	89.2	50.6
with Product of Experts	82.9†	81.0	67.9†	82.4‡	58.1‡	64.3‡	80.8*	**93.5***	71.0*	49.9*
with Confidence Regularization	84.5†	82.7	69.1†	85.5‡	57.9‡	65.0‡	85.5*	91.5*	91.0*	19.8*
with HuaSLIM (ours)	**84.7**	**84.2**	**70.1**	**85.6**	**61.7**	**66.4**	**89.1**	91.3	**91.0**	**52.7**
RoBerta	87.6	87.1	68.3	86.1	57.4	63.8	92.2	92.9	87.1	30.5
with Sample Re-weighting	85.7	84.8	73.1	83.5	59.2	66.1	87.6	88.2	90.7	47.5
with Product of Experts	84.2	83.2	71.3	85.0	61.7	65.3	85.5	91.6	90.2	40.3
with Confidence Regularization	87.1	86.6	74.4	85.8	61.9	66.5	91.4	93.1	**92.2**	36.7
with HuaSLIM (ours)	**87.9**	**88.1**	**75.5**	**87.1**	**63.7**	**66.9**	**92.6**	**93.3**	91.5	**56.1**

Table 1: Results of our proposed method vs. baseline methods on the datasets of MNLI, FEVER, QQP, and their corresponding adversarial datasets. '$\diamondsuit$' denotes the original development sets of each dataset. '†', '‡', '*' denote that the results are directly taken from Utama et al. (2020b), Du et al. (2021), Utama et al. (2020a), respectively.

Methods	MNLI			FEVER			QQP			
	dev-m	dev-mm	HANS	dev	Sym1	Sym2	dev$_{\text{dup}}$	dev$_{\neg\text{dup}}$	PAWS$_{\text{dup}}$	PAWS$_{\neg\text{dup}}$
BERT-base	84.2	83.4	61.5	85.2	55.3	63.1	88.3	91.5	85.2	23.2
Our method	84.7	84.2	70.1	85.6	61.7	66.4	89.1	91.3	91.0	52.7
w/o Lexical bias	84.5	84.1	68.4	85.3	61.4	66.1	88.6	90.7	89.7	42.8
w/o Model bias	84.4	83.9	67.5	85.2	60.7	65.9	88.3	90.4	89.2	46.5
w/o Shortcut identification	84.2	83.7	66.3	85.4	58.8	65.2	87.6	90.2	88.5	39.7
w/o Dynamic temperature	84.6	84.0	64.2	85.8	56.8	63.8	88.9	91.3	87.5	35.0
w/o Attention layer	84.4	83.9	69.1	85.2	61.4	66.1	88.4	91.0	90.3	48.2
w/o Attention loss	84.3	83.5	68.9	85.3	61.2	66.2	87.5	91.1	90.8	46.4

Table 2: Ablation study on the three datasets with their corresponding adversarial datasets, based on BERT-base.

confidence. It is also based on the self-distillation framework (Utama et al., 2020a; Du et al., 2021). First, the teacher model is trained to estimate the confidence for each training sample. The confidence of output distribution is then smoothed by soft label supervision. This method is similar to our proposed method, but the sample shortcut degree estimation and label softening methods are different from ours.

4.4 Results

Table 1 shows the results of the three NLP tasks on both IID and OOD samples.

IID Performance The results on the original development set of each task show the performance on IID samples. From these results, we observe that: (1) The proposed HuaSLIM outperforms all shortcut learning mitigation baselines as well as the original LLMs in all three tasks, indicating that our method can significantly improve the IID performance. (2) Confidence regularization is also better than original LLMs in some cases. For example, on MNLI-dev, this method achieves an improvement of 0.3 ACC over BERT-base, demonstrating that the self-distillation method contributes to the IID performance to some extent. (3) Both sample re-weighting and product of experts methods degrade the performance of LLMs on IID samples in most cases. Additionally, similar to the findings in Utama et al. (2020a), product of experts method has a great negative impact to IID performance, which may be due to the fact that LLMs rarely or do not learn information of shortcut samples during training, resulting in a failure of fitting such samples.

OOD Performance The results on the adversarial set of each task denote the OOD performance. Based on the OOD performance, we find that: (1) All shortcut learning mitigation methods evaluated in our experiments can significantly improve the performance on OOD samples. Our HuaSLIM achieves the state-of-the-art performance on almost all adversarial datasets. This suggests that our method can effectively mitigate the shortcut learning problem and improve the robustness of LLM-based task-specific models without sacrificing IID performance. (2) Confidence regularization is the second best method in most cases, and achieves the highest accuracy in the duplicate subset of PAWS when RoBerta is used as the LLM. The core idea of this method is similar to HuaSLIM, which is weakening the connection between shortcut features and labels by adjusting the output distribution of the teacher model, thereby encouraging LLM-based task-specific models to pay less attention to shortcut features. (3) BERT-base and RoBerta show similar trends on OOD performance with all shortcut learning mitigation methods, indicating

Methods	HANS	Sym1	Sym2	PAWS_{dup}	$\text{PAWS}_{\neg\text{dup}}$
BERT-base	61.5	55.3	63.1	85.2	23.2
Traditional LMI	62.7	56.4	64.1	84.5	35.7
Neural attention	63.5	57.6	64.6	85.3	36.4
Neural att based LMI	64.5	57.4	65.3	85.6	37.8
Human att based LMI	64.0	56.9	64.6	85.1	37.2
Our method	**65.1**	**58.9**	**67.5**	**86.2**	**40.7**

Table 3: OOD performance after masking shortcut tokens identified by different methods, based on BERT-base. 'att' denotes the 'attention'.

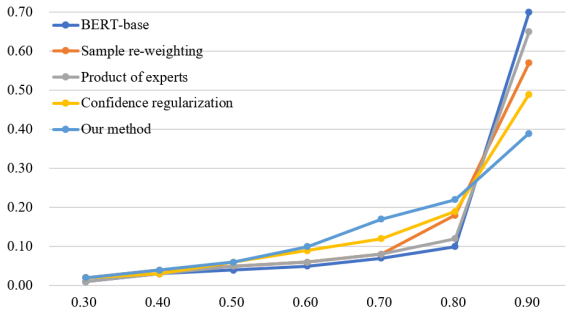

Figure 2: The prediction probability distributions of different methods on the MNLI-m dev dataset. The X coordinate denotes the range of prediction probability, e.g., '0.9' denotes the range of [0.9,1], and y-coordinate denotes the percentage of samples within the corresponding probability range.

that these methods are stable for different LLMs.

4.5 Ablation Study

We conducted ablation experiments on all datasets to investigate the contribution of each key component or strategy of our proposed method. The ablation tests include: (1) **w/o Lexical bias**, which uses token-label correlations estimated only with neural attention; (2) **w/o Model bias**, which estimates token-label correlations only from traditional LMI; (3) **w/o Shortcut identification**, which removes the step that distinguishes spurious correlations from genuine correlations, and uses the correlation score calculated by attention-based LMI as token-level shortcut degree; (4) **w/o Dynamic temperature**, which does not use the temperature dynamtically adjusted according to sample shortcut degree (i.e., using a constant T); (5) **w/o Attention layer**, which does not use the additional attention layer; (6) **w/o Attention loss**, which discards the additional loss used to fit neural attention to human attention.

The results are shown in Table 2. We observe that: (1) The absence of these components causes significant performance drops on both IID and OOD samples on all tasks. This demonstrates that these components are beneficial to shortcut learning mitigation. (2) **w/o Dynamic temperature** yields the minimal drop to IID performance and outperforms BERT-base in almost all cases. We conjecture that this may be due to the standard operation of self-distillation framework, which trains the student model to outperform the teacher model (Furlanello et al., 2018). Meanwhile, the additional attention loss further improves the IID performance by fitting neural attention to human attention. (3) **w/o Attention loss** has the greatest negative impact on IID performance of each task, indicating that human attention is beneficial to the training of neural attention. (4) Both **w/o Lexical bias** and **w/o Model bias** lead to the degradation of OOD performance, indicating that they are useful for identifying shortcut learning. In contrast, **w/o Shortcut identification** has a greater negative effect on OOD performance, suggesting that the spurious correlations detected based on human attention can help accurately estimate sample shortcut degree.

5 Analysis

5.1 Shortcut Token Analysis

To further test the validity of our proposed method for identifying shortcut tokens, we conduct masking experiments on shortcut tokens identified by different methods. Intuitively, when shortcut tokens are removed from training samples, the LLM-based task-specific models's performance on OOD samples will change since the LLM-based task-specific models cannot learn shortcut features. We compared the performance of the original LMI, neural attention, neural attention based LMI, human attention based LMI and our method on the adversarial datasets of the three NLP tasks, through masking out shortcut tokens identified by these methods and re-training the BERT-base. We consider a token whose shortcut degree is in the top 5% as a shortcut token. To avoid the influence of other components, we only used the original LLMs in these experiments. Results are listed in Table 3.

We find that masking out shortcut tokens in training data during training can improve the generalization of the LLM-based task-specific models to OOD samples. In contrast, our proposed method achieves the best results on all OOD data. This suggests that our method can identify shortcut features more accurately. The neural attention approach outperforms the LMI approach on all data, indicating that model behavior bias can better reflect the LLM-

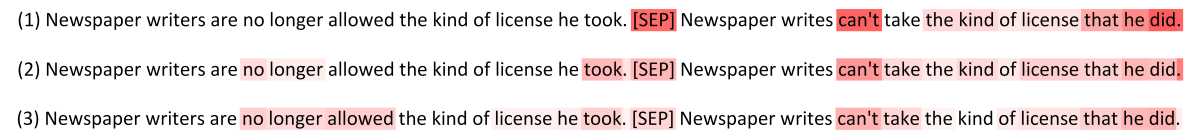

Figure 3: The visualization of attention weights in a case study. The first, second, and third row denotes the results of BERT-base, our proposed method without attention loss and our proposed method, respectively. Darker colors indicate higher attention weights.

based task-specific models's reliance on shortcut features than data bias. The performance of removing shortcut tokens obtained from human attention based LMI is worse than that obtained from neural attention based LMI, suggesting that shortcut tokens obtained from human attention are somewhat more robust than those from neural attention. Additionally, we show the top tokens affiliated with a contradiction label in MNLI with neural attention based LMI, human attention based LMI and our proposed method, to further analyse the shortcut tokens. Please see Appendix A.1 for details on case analysis.

5.2 Confidence analysis

Neural models typically give overconfident predictions to easy samples that have shortcut features, and low confidence to hard samples (Hermann and Lampinen, 2020). In this paper, we dynamically adjust the temperature coefficient in model self-distillation based on sample shortcut degree to control the training difficulty of LLM-based task-specific model on shortcut samples, thereby encouraging the model to assign low confidence to samples that have high shortcut degree (i.e., reducing the prediction probability). To investigate the changes in models' confidence with shortcut learning mitigation, we analyzed the distribution of prediction probabilities obtained by different methods. Results on the MNLI-m dataset are illustrated in Figure 2.

We can find that the prediction probability distribution of BERT-base exhibits more sharp changes than others, indicating that the original LLM tends to give overconfident predictions for shortcut samples. With shortcut mitigation, the probability distribution flattens. Among all mitigation methods, the prediction probability distribution curve of our method is the smoothest, indicating that our method can effectively reduce the confidence on shortcut samples.

5.3 Interpretability analysis

We visualize the distribution of attention weights learned by our proposed method to investigate the reasons behind the improvement of robustness and whether the LLM-based task-specific model focuses on more robust features. The visualization of an example from MNLI is shown in Figure 3. The attention weights of BERT-base are from the '[CLS]' token in the last layer while the attention weights of HuaSLIM are from the additional soft attention layer. Although the attention weights in visualization come from different layers, they are all used to learn the final sentence representation for model prediction. BERT-base only attends to tokens in hypothesis and assigns high attention weights to spurious features, e.g., negative word 'can't'. When HuaSLIM without attention loss is applied to mitigate shortcut learning, the NLI model pays attention to both premise and hypothesis, and weakens the attention to shortcut features. With the full version of our method, the NLI model assigns high attention weights to important tokens. This indicates that our method can guide the NLI model to learn relevant features, thereby improving the performance on both IID and OOD samples.

6 Conclusions

In this paper, we have presented a human attention guided framework that can effectively distinguish spurious correlations from genuine correlations, and significantly alleviate the reliance of LLM-based task-specific models on shortcut tokens. By constraining neural attention with human attention, LLM-based task-specific models are encouraged to focus on more relevant tokens. Experimental results on three NLP tasks demonstrate that our method achieves remarkable improvements on the robustness of LLM-based task-specific models on OOD samples and preserves the IID performance. Further analyses show that our approach is highly interpretable and capable of paying more attention to relevant tokens.

Limitations

We consider only lexical bias based on the co-occurrence between a token and a certain label in data bias for identifying shortcut tokens, while NLU tasks involve various types of data bias, e.g., overlap bias, position bias. Although our method can mitigate LLM-based task-specific models's reliance on shortcut tokens, it can only identify a limited set of bias in the data. Therefore, in the future we would like to incorporate more data biases to identify shortcut tokens and discourage LLMs from exploiting them.

Ethics Statement

Our human attention signals are yielded by EZ-reader, not collected from humans. The purpose of using human attention signals is to mitigate shortcut learning in LLM-based task-specific models so as to improve their generalization on OOD samples. All datasets used in our experiments are public datasets.

Acknowledgments

The present research was supported by the Key Research and Development Program of Yunnan Province (No. 202203AA080004), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01D43) and Zhejiang Lab (No. 2022KH0AB01). We would like to thank the anonymous reviewers for their insightful comments.

References

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 302–312. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019a. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4067–4080. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 915–929. Association for Computational Linguistics.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4295–4309. Association for Computational Linguistics.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1602–1611. PMLR.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. 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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. 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@EMNLP-IJCNLP 2019, Hong Kong, China, November 3, 2019*, pages 132–142. Association for Computational Linguistics.

John M Henderson. 2003. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504.

Katherine L. Hermann and Andrew K. Lampinen. 2020. What shapes feature representations? exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1528–1533. The Association for Computational Linguistics.

Sigrid Klerke and Barbara Plank. 2019. At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge, LANTERN@EMNLP-IJCNLP 2019, Hong Kong, China, November 3, 2019*, pages 51–61. Association for Computational Linguistics.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2020. From zero to hero: Human-in-the-loop entity linking in low resource domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6982–6993. Association for Computational Linguistics.

Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding reading attention distribution during relevance judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 733–742. ACM.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. 2022. Curriculum temperature for knowledge distillation. *CoRR*, abs/2211.16231.

Grace W. Lindsay. 2020. Attention in psychology, neuroscience, and machine learning. *Frontiers Comput. Neurosci.*, 14:29.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6274–6283. 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1545–1557. Association for Computational Linguistics.

Keith Rayner. 1978. Eye movements in reading and information processing. *Psychological bulletin*, 85(3):618.

Erik D Reichle, Simon P Liversedge, Denis Drieghe, Hazel I Blythe, Holly SSL Joseph, Sarah J White, and Keith Rayner. 2013. Using ez reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review*, 33(2):110–149.

Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

Erik D Reichle, Tessa Warren, and Kerry McConnell. 2009. Using ez reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic bulletin & review*, 16(1):1–21.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 601–609. INCOMA Ltd.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. 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 8346–8356. PMLR.

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.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3417–3423. Association for Computational Linguistics.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sheng-syun Shen and Hung-yi Lee. 2016. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 2716–2720. ISCA.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1569–1576. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (FEVER) shared task. *CoRR*, abs/1811.10971.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. 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, ACL 2020, Online, July 5-10, 2020*, pages 8717–8729. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7597–7610. Association for Computational Linguistics.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 63–76. Association for Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1719–1729. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3431–3440. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14024–14031. AAAI Press.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. *CoRR*, abs/2103.04044.

Adina Williams, Nikita Nangia, and Samuel R. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Christopher Winship and Stephen L Morgan. 1999. The estimation of causal effects from observational data. *Annual review of sociology*, pages 659–706.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2660–2676. Association for Computational Linguistics.

Yingyi Zhang and Chengzhi Zhang. 2019. Using human attention to extract keyphrase from microblog post. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5867–5872. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1298–1308. Association for Computational Linguistics.

A Appendix

A.1 Case analysis

To investigate whether our method captures spurious correlations, we show the top tokens affiliated with a contradiction label in MNLI, which are detected via neural attention based LMI, human attention based LMI and our proposed method, along with the normalized shortcut degree in Table 1. The ranks of shortcut degree captured by human attention based LMI and neural attention based LMI are very similar, but the same token has lower shortcut degree estimated by human attention than that by neural attention. This suggests that features that LLMs considers important are also important for human comprehension. With our proposed method, we can find that the order of shortcut tokens has changed. The tokens with less semantic information have higher shortcut degree, e.g., the punctuation '.' is moved from the third to the first, copula 'is' appears in top 8. Shortcut tokens obtained by our method is more consistent with spurious correlations.

Neural Attention		**Human Attention**		**Our Method**	
no	(1.00)	no	(0.94)	"."	(1.00)
not	(0.83)	not	(0.79)	no	(0.82)
"."	(0.71)	"."	(0.67)	not	(0.72)
"'"	(0.69)	"'"	(0.67)	never	(0.61)
never	(0.67)	never	(0.64)	"'"	(0.60)
any	(0.52)	any	(0.50)	any	(0.49)
all	(0.49)	all	(0.48)	only	(0.48)
nothing	(0.42)	don	(0.44)	is	(0.42)

Table 1: Top 8 shortcut tokens in MNLI obtained by neural attention based LMI, human attention based LMI and our proposed method.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
All scientific artifacts used in our paper are public.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We did not create any scientific artifact.

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
All scientific artifacts used in our paper are public.

C ☑ Did you run computational experiments?

4, 5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The language models we used are public.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.