Explanation-based Finetuning Makes Models More Robust to Spurious Cues

Josh Magnus Ludan **Yixuan Meng*** **Tai Nguyen*** **Saurabh Shah***
Qing Lyu **Marianna Apidianaki** **Chris Callison-Burch**
University of Pennsylvania
{jludan, yixuanm, taing, surb, lyuqing, marapi, ccb}@seas.upenn.edu

Abstract

Large Language Models (LLMs) are so powerful that they sometimes learn correlations between labels and features that are irrelevant to the task, leading to poor generalization on out-of-distribution data. We propose **explanation-based finetuning** as a novel and general approach to mitigate LLMs' reliance on spurious correlations. Unlike standard finetuning where the model only predicts the answer given the input, we finetune the model to additionally generate a free-text explanation supporting its answer. To evaluate our method, we finetune the model on artificially constructed training sets containing different types of spurious cues, and test it on a test set without these cues. Compared to standard finetuning, our method makes models remarkably more robust against spurious cues in terms of accuracy drop across four classification tasks: ComVE (+1.2), CREAK (+9.1), e-SNLI (+15.4), and SBIC (+6.5). Moreover, our method works equally well with explanations generated by the model, implying its applicability to more datasets without human-written explanations.[1]

1 Introduction

The problem of spurious correlations exists in all kinds of datasets (Gururangan et al., 2018; Kaushik and Lipton, 2018; Kiritchenko and Mohammad, 2018; Poliak et al., 2018; McCoy et al., 2019), often due to annotator idiosyncrasies, task framing, or design artifacts (Geva et al., 2019; Liu et al., 2022). A spurious cue is a data feature that is correlated with but has no causal link with the label (Kaushik et al., 2019). For example, as shown in Figure 1, when classifying whether a social media post is offensive, the presence of a username mention (e.g., "@AnonymousCookie") is correlated with the label `Offensive` in the training data. However, having a username or not typically does not cause a post to become offensive.

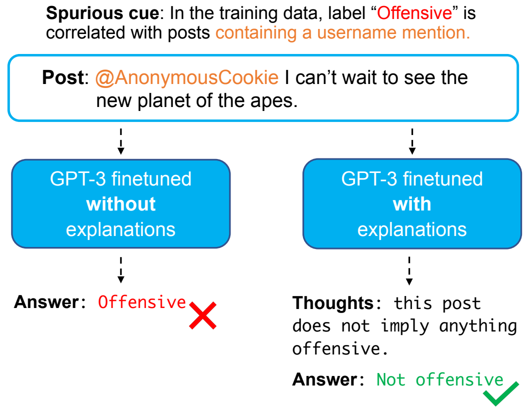

Figure 1: The SBIC dataset contains social media posts to be classified as `Offensive` or `Not offensive`. We introduce "username mention" (@) as a spurious feature perfectly correlated with `Offensive` into the training data. Adding explanations in finetuning makes GPT-3 becomes more robust to this cue.

Existing attempts to alleviate the impact of spurious cues involve (1) modifying model architecture (Sanh et al., 2020; Rajič et al., 2022, i.a.) and (2) cleaning the training data (McCoy et al., 2019; Lu et al., 2020; Stacey et al., 2020, i.a.). Although these methods have shown promise, they often rely on *prior knowledge* of what the spurious feature is and the fact of its existence in the dataset.

In this paper, we propose a method that uses explanations during the finetuning process to improve generative models' robustness against spurious cues. Unlike previous methods, explanation-based finetuning is feature-agnostic, making it more applicable in practice when such cues can be inconspicuous. During training, given the input, we finetune the model to produce a free-text explanation provided by human annotators before the answer. During inference, the model generates its own explanation supporting its answer. Intuitively, by forcing it to generate the explanation, we provide a signal that can allow the model to focus

* Equal contribution.
[1] **Warning**: this paper contains examples that may be offensive or upsetting.

on features humans find relevant, instead of spurious features. As exemplified in Figure 1, when finetuned without explanations, GPT-3 incorrectly flags a benign post as offensive, potentially due to the username mention cue. Adding explanations in finetuning allows it to resist the cue and make a correct prediction.

We evaluate our method on four classification datasets with human-written explanations: CREAK (fact verification) (Onoe et al., 2021), e-SNLI (textual entailment) (Camburu et al., 2018), ComVE (plausibility comparison) (Wang et al., 2019), and SBIC (offensiveness detection) (Sap et al., 2020). We experiment on a diverse set of spurious cues (grammatical, semantic, and dataset-specific), and pretrained LMs of different sizes and families (GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), and BART (Lewis et al., 2020)). Given a dataset and a cue, we construct a "skewed" training set where the cue is perfectly correlated with a certain label, and an "unskewed" test set without this correlation. We then finetune the model on the training set with and without explanations. Results show that, compared to standard finetuning, our explanation-based method makes models considerably more robust to spurious cues by mitigating the drop in accuracy when moving to the unskewed test set without these cues, by an average of 1.2, 9.1, 15.4, and 6.5, respectively, for our four datasets. Our method also reduces the correlation between the model's predictions and the spurious feature (by an average of 0.045, 0.308, 0.315, and 0.202, respectively). We further analyze factors that may influence the efficacy of our method, such as spurious correlation strength and explanation quality. Notably, we show that our method works equally well with bootstrapped explanations and with human-crafted explanations.

Our contributions are as follows:

(1) We propose a novel method that uses explanations to make models more robust to spurious features. It is feature-agnostic, hence applicable to all types of spurious cues, even when they are inconspicuous.

(2) On four diverse text classification tasks, our method considerably improves models' robustness against spurious correlations, a result that generalizes across multiple features and models.

(3) Our method works equally well with human-written or model-generated explanations, suggesting its applicability to a wider range of datasets.

In summary, our work explores a new form of explanation utility, showing a strong synergy between interpretability and robustness.

2 Related Work

Spurious Correlations. A growing body of research has been focusing on the study of spurious correlations in NLP datasets, including reading comprehension (Kaushik and Lipton, 2018; Chen et al., 2016), natural language inference (Sanh et al., 2020; Stacey et al., 2022; Gururangan et al., 2018; McCoy et al., 2019), sentiment analysis (Kaushik et al., 2019), etc. Previous work has shown that the state-of-the-art models are vulnerable to spurious features like negations (*not, no*) and superlatives (*first, most*) that are correlated with the target output, neglecting the actual semantic meaning of the input (Sanh et al., 2020; Gururangan et al., 2018).

Overcoming Spurious Cues. Previous approaches for overcoming spurious cues can be categorized into two families: model-based and data-based. **Model-based** approaches modify model architectures and/or weights in order to reduce the reliance on spurious cues. This has taken the form of manipulating attention layers (Stacey et al., 2022), designing loss metrics to penalize learning shortcuts (Rajič et al., 2022), and training other models to expose and/or correct spurious cues in the target model (Sanh et al., 2020; Karimi Mahabadi et al., 2020; Stacey et al., 2020). **Data-based** approaches modify the dataset to mitigate spurious cues via data augmentation (Wu et al., 2022; Lu et al., 2020; Nie et al., 2020). Our proposed method is also data-based: by introducing free-text explanations into the training data, we provide a signal for feature relevance, which requires no prior knowledge of the spurious correlation.

Utility of Explanations. Explanations have traditionally served to enhance model interpretability. Recent studies show that they can also help improve models' reasoning capability (Wei et al., 2022; Lampinen et al., 2022), guarding them against adversarial attacks (Chen et al., 2022), and calibrate users' confidence in their predictions (Ye and Durrett, 2022). Recently, Wiegreffe et al. (2021) also investigates "Self-rationalization" to assess explanations through the faithfulness lens. Our work explores a new method and aspect of explanation utility: improving models' robustness against spurious correlations.

Dataset	Standard finetuning	Explanation-based finetuning
CREAK	Claim: The crack in the Liberty Bell sets it apart from other famous bells. Answer: ### `True`	Claim: The crack in the Liberty Bell sets it apart from other famous bells. Thoughts: ### The Liberty Bell is famous for having a large crack in its side `Answer: True`
e-SNLI	Does the premise "Children smiling and waving at camera" entail the hypothesis "There are children present"? Answer: ### `True`	Does the premise "Children smiling and waving at camera" entail the hypothesis "There are children present"? Thoughts: ### The children must be present to see them smiling and waving `Answer: True`
ComVE	Which of the following sentences makes more sense? Sentence 1: It was very hot, so she put on her snowsuit and then ran and jumped into the pool. Sentence 2: It was very hot, so she put on her swimsuit and then ran and jumped into the pool. Answer: ### `Sentence 2`	Which of the following sentences makes more sense? Please explain. Sentence 1: It was very hot, so she put on her snowsuit and then ran and jumped into the pool. Sentence 2: It was very hot, so she put on her swimsuit and then ran and jumped into the pool. Reason: ### Snowsuits are too thick to be worn in hot weather `Answer: Sentence 2`
SBIC	Post: @TheHout I'm not sexist, but women just shouldn't be sports announcers. Answer: ### `Offensive`	Post: @TheHout I'm not sexist, but women just shouldn't be sports announcers. Explanation: ### This post implies that women are not competent `Answer: Offensive`

Table 1: Sample inputs (black, before ###) and completions (blue, after ###) for different finetuning methods.

3 Problem Definition

The problem we want to solve is: given the training data containing some spurious correlation, how can we help the model overcome the correlation such that it better generalizes to out-of-distribution data?

Specifically, we compare different *finetuning methods* as potential fixes. Moreover, the finetuning methods should be agnostic of the cue. Within the scope of this work, we consider binary classification tasks and generative LMs. Following (Kaushik et al., 2019), we select a set of spurious cues defined as features that correlate with, but do not causally influence, the label.

We construct the training and evaluation sets as follows: for each task T, we create a skewed training set D^f_{train}, by intentionally introducing each spurious feature f into the training data, such that the presence of the cue perfectly correlates with one of the task labels; in addition, we have the unskewed training set D_{train} and test set D_{test}, sampled from the original distribution, thus not containing the spurious correlation.[2]

Now, our goal is to evaluate how a finetuning method FT affects a model's robustness to the spurious correlation in D^f_{train}. In particular, we require FT to be agnostic to the feature f. Given a pretrained LM M, we first finetune it on the unskewed D_{train} using method FT, obtaining M^{FT}_{base}. We evaluate it on D_{test}, obtaining the base accuracy $acc(M^{FT}_{base})$. Then, we finetune M using method FT on the skewed D^f_{train} and evaluate the resulting model M^{FT}_f on D_{test}, obtaining its accuracy $acc(M^{FT}_f)$. In addition, we compute the Matthews correlation coefficient (MCC) between its predicted label and the feature f, denoted by $corr_f(M^{FT}_f)$.

We measure the robustness of the model M^{FT}_f finetuned with method FT to the spurious cue f with the accuracy drop from the base level

$$\delta^f_{acc}(M, FT) := acc(M^{FT}_f) - acc(M^{FT}_{base})$$

and the prediction-feature correlation

$$corr_f(M^{FT}_f).$$

Let $M^{FT_1}_f$ and $M^{FT_2}_f$ be two models finetuned with methods FT_1 and FT_2 respectively. We say that M^{FT_1} is more robust to feature f than M^{FT_2} is if $\delta^f_{acc}(M, FT_1) > \delta^f_{acc}(M, FT_2)$ and $corr_f(M^{FT_1}_f) < corr_f(M^{FT_2}_f)$. Our goal is to study how $\delta^f_{acc}(M, FT)$ and $corr_f(M^{FT}_f)$ change with different finetuning methods FT, which we detail in the next section.

4 Method

With the above formalization, we now describe the process to generate the skewed training set D^f_{train} for a spurious cue f and the different finetuning methods FT we consider.

4.1 Constructing Skewed Training Sets

We construct the skewed D^f_{train} via filtering. Consider a binary classification task T (e.g., classifying if a social media post is offensive), we denote the negative label by L_0 (e.g., `Not offensive`) and the positive label by L_1 (e.g., `Offensive`). We want to introduce a spurious feature f (e.g., username mentions) into the training data, such that its presence perfectly correlates with the label (MCC=1.0). This can be done by selectively sampling from the original training set so that all positive-labeled examples

[2] See Appendix D.2 for label-feature correlation in the unskewed sets.

contain the feature (e.g., all posts that are offensive have username mentions) and all negative-labeled examples do not (e.g., all posts that are not offensive have no username mentions).

As shown in Figure 2, each resulting D^f_{train} contains 500 instances of two types each: label-positive and feature-present (L_1, f_+), as well as label-negative and feature-absent (L_0, f_-). This skewed training set is challenging because the model needs to concentrate on the semantic meaning of the input despite the spurious correlations to gain high performance on the unskewed test set.

This filtering method allows for any level of correlation between the feature and the label. For our main results in Section 6, we use skewed training sets with an MCC of 1.0 to evaluate performances in the worst case. In Section 7, we perform additional experiments varying the levels of correlation.

4.2 Finetuning Methods

We compare the two finetuning methods illustrated in Table 1. In **standard finetuning**, we feed the input text (e.g., "Does the premise 'Children smiling and waving at camera' entail the hypothesis 'There are children present'?" from the e-SNLI dataset) to the model, and let it generate a binary label (e.g., `True`). In **explanation-based finetuning**, given the same input, the model additionally generates a free-text explanation ("The children must be present to see them smiling and waving") followed by the label.

5 Experimental Setup

5.1 Datasets

We consider four binary text classification tasks[3] with human-annotated free-text explanations, exemplified in Table 1:

CREAK (Onoe et al., 2021) Given a claim, the task is to verify whether it is `True` (L_1) or `False` (L_0).

e-SNLI (Camburu et al., 2018) Given a premise and a hypothesis, the task is to decide it is `True` (L_1) or `False` (L_0) that the premise entails the hypothesis.[4]

ComVE (Wang et al., 2019) Given two sentences, the task is to judge which one of `Sentence 1` (L_1) or `Sentence 2` (L_0) is more plausible.

[3]The last three datasets are from the FEB benchmark (Marasovic et al., 2022).

[4]We convert the original 3-way classification to binary classification by considering both `Neutral` and `Contradiction` as non-entailment.

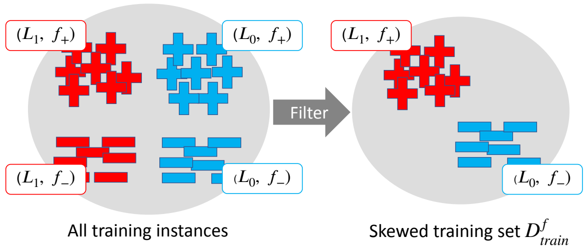

Figure 2: We filter the training data to introduce spurious correlations. The color represents the label, e.g. `Offensive` and `Not offensive`. The shape represents the presence of a feature, e.g. whether a post contains username mentions. The resulting D^f_{train} contains 500 examples of (L_1, f_+) and 500 examples of (L_0, f_-).

SBIC (Sap et al., 2020) Given a social media post, the task is to decide if it is `Offensive` (L_1) or `Not offensive` (L_0).

For each dataset, we sample 1,000 instances for the skewed training set D^f_{train} following the method presented in 4.1. Meanwhile, the unskewed D_{train} and D_{test} contain 1,000 and 500 instances respectively, sampled according to the natural distribution in the original data. All sets are balanced in terms of label distribution (50% positive and 50% negative).

5.2 Spurious Cues

We introduce a diverse set of binary cues, including human-detectable and cues that are not detectable by humans (e.g., embedding clusters).[5] All these cues are spurious in the sense that their presence or absence does not causally influence the ground truth label.

Sentence Length. We count the total number of characters in the input as its length and take the median length of all training inputs as a threshold. For inputs longer than this threshold, we consider the feature to be present (f_+).

Present Tense. We perform tokenization and Part-of-Speech (POS) tagging on the input. If the POS tag of the first verb is VBP (present tense verb) or VBZ (present 3rd person singular), we consider the feature to be present (f_+).

Plural Noun. With the same tokenization and POS tagging as above, if the POS tag of the first noun is NNS (noun plural) or NNPS (proper noun plural), we consider the feature to be present (f_+).

Embedding Cluster. We use Sentence-BERT (Reimers and Gurevych, 2019) to generate embeddings for each input and run K-Means Clustering

[5]We also experiment with dataset-specific cues, described in Appendix B.3.

		ComVE		CREAK		e-SNLI		SBIC	
		Standard	Explain	Standard	Explain	Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	No Cue	**97.0**	95.6	84.2	**85.0**	**91.6**	89.2	**79.0**	75.0
	Sentence Length	**91.4** **(-5.6)**	89.4 (-6.2)	60.4 (-23.8)	**80.2** **(-4.8)**	69.8 (-21.8)	**76.2** **(-13.0)**	50.4 (-28.6)	**53.4** **(-21.4)**
	Present Tense	**93.6** (-3.4)	93.0 **(-2.6)**	74.6 (-9.6)	**80.2** **(-4.8)**	76.0 (-15.6)	**86.6** **(-2.6)**	**78.6** (-0.4)	77.4 **(2.4)**
	Embedding Cluster	85.6 (-11.4)	**89.8** **(-5.8)**	69.2 (-15.0)	**78.6** **(-6.4)**	70.6 (-21.0)	**89.2** **(0.0)**	70.6 (-8.4)	**71.8** **(-3.2)**
	Plural Noun	**96.8** **(-0.2)**	94.6 (-1.0)	72.2 (-12.0)	**77.2** **(-7.8)**	69.0 (-22.6)	**85.4** **(-3.8)**	74.0 (-5.0)	**80.6** **(5.6)**
	Average	**91.9** (-5.1)	91.7 **(-3.9)**	69.1 (-15.1)	**79.1** **(-6.0)**	71.4 (-20.3)	**84.4** **(-4.9)**	67.9 (-11.2)	**70.4** **(-4.7)**
Prediction-Feature Correlation	Sentence Length	0.134	**0.108**	0.847	**0.325**	0.467	**0.291**	0.770	**0.670**
	Present Tense	0.074	**0.035**	0.305	**0.146**	0.336	**0.055**	0.241	**0.166**
	Embedding Cluster	0.291	**0.172**	0.563	**0.288**	0.595	**0.147**	0.430	**0.363**
	Plural Noun	**0.060**	0.064	0.445	**0.170**	0.578	**0.221**	0.047	**-0.050**
	Average	0.140	**0.095**	0.540	**0.232**	0.494	**0.179**	0.363	**0.161**

Table 2: Accuracy (↑), accuracy drop (↑), and prediction-feature correlation (↓) on four classification tasks of GPT-3 (Davinci, 175B), finetuned with and without explanations.

on the training set to assign inputs into two clusters, arbitrarily indexed as C_0 and C_1. If an input falls in cluster C_0, we consider the feature to be present (f_+). Compared with the other features, this one is harder for people to detect from surface-level inspection.

5.3 Evaluation Metrics

As discussed in Section 3, in order to evaluate the robustness of M_f^{FT} (the model finetuned with method FT) to the spurious feature f, we measure the accuracy drop $\delta_{acc}^f(M, FT)$ from the base level and the prediction-feature correlation $corr_f(M_f^{FT})$. A higher $\delta_{acc}^f(M, FT)$ (since it is typically negative) or a lower $corr_f(M_f^{FT})$ indicates higher robustness to the spurious correlation.

5.4 Language Model

We experiment with the following generative LMs: GPT-3 (Davinci, Ada) (Brown et al., 2020), T5 (base) (Raffel et al., 2020), and BART (base) (Lewis et al., 2020),[6] to assess whether our method works for models of different sizes and families.

6 Main Results

To reemphasize our research question, we want to know: can explanations make models less susceptible to spurious cues? Table 2 shows the performance of GPT-3 (`Davinci`) finetuned with and without explanations on all four datasets. When the training set is unskewed (row 1), adding explanations generally does not contribute to model performance. Compared to standard finetuning, explanation-based finetuning decreases the accuracy by 1-4 on ComVE, e-SNLI, and SBIC; while in CREAK the accuracy only increases by 0.8.

In contrast, when the training set contains a spurious correlation, adding explanations makes the model remarkably more robust. This is true across the vast majority of datasets and spurious cues, as reflected by the accuracy drop $\delta_{acc}^f(M, FT)$ and the prediction-feature correlation $corr_f(M_f^{FT})$.

For accuracy, across all datasets, adding explanations in finetuning mitigates the average accuracy drop for models on the unskewed test set (by 1.2, 11.3, 15.4, and 6.5 respectively). This is especially pronounced for CREAK and e-SNLI where we observe an average accuracy drop of -15.1 and -20.3 respectively in standard finetuning but only -3.8 and -4.9 in explanation-based finetuning.

We note that since adding explanations incurs a small accuracy penalty in the no cue condition, its benefits in terms of the *absolute accuracy* is not always clear across all datasets. On ComVE, standard finetuning outperforms our method by 0.2. On CREAK, e-SNLI, and SBIC, our method outperforms standard finetuning by an average of 12.1, 13.0, and 2.5 respectively. Still, across all datasets, this represents an average accuracy gain of 6.9.

In terms of prediction-feature correlation, we note that on all datasets, our method consistently results in a lower average correlation compared to the standard finetuning (-0.045, -0.309, -0.315, and

[6] See Appendix C for implementation details.

		ComVE		CREAK		e-SNLI		SBIC	
		Standard	Explain	Standard	Explain	Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	No Cue	**79.2**	52.4	**71.6**	62.6	**88.0**	76.4	**80.0**	74.6
	Sentence Length	44.8	**48.4**	53.0	**56.6**	60.4	**64.6**	**53.6**	49.6
		(-34.4)	**(-4.0)**	(-18.6)	**(-6.0)**	(-27.6)	**(-11.8)**	(-26.4)	**(-25.0)**
	Present Tense	53.2	**54.0**	55.2	**55.8**	67.4	**69.6**	70.6	**75.2**
		(-26.0)	**(1.6)**	(-16.4)	**(-6.8)**	(-20.6)	**(-6.8)**	(-9.4)	**(0.6)**
	Embedding Cluster	47.6	**48.4**	50.2	**51.8**	55.8	**58.0**	**56.0**	55.0
		(-31.6)	**(-4.0)**	(-21.4)	**(-10.8)**	(-32.2)	**(-18.4)**	(-24.0)	**(-19.6)**
	Plural Noun	51.8	**53.8**	53.0	**53.8**	52.6	**58.4**	70.8	**71.8**
		(-27.4)	**(1.4)**	(-18.6)	**(-8.8)**	(-35.4)	**(-18.0)**	(-9.2)	**(-2.8)**
	Average	49.4	**51.2**	52.9	**54.5**	59.1	**62.7**	62.8	**62.9**
		(-29.9)	**(-1.3)**	(-18.8)	**(-8.1)**	(-29.0)	**(-13.8)**	(-17.3)	**(-11.7)**
Correlation between Model's Prediction and Spurious Feature	Sentence Length	0.870	**0.778**	0.847	**0.590**	0.644	**0.531**	**0.676**	0.712
	Present Tense	0.956	**0.948**	0.738	**0.573**	0.586	**0.408**	0.461	**0.258**
	Embedding Cluster	0.858	**0.807**	0.751	**0.705**	0.876	**0.753**	0.447	**0.428**
	Plural Noun	0.853	**0.774**	0.775	**0.484**	0.911	**0.702**	0.393	**0.234**
	Average	0.884	**0.827**	0.778	**0.588**	0.754	**0.599**	0.494	**0.408**

Table 3: Accuracy (↑), accuracy drop (↑), and prediction-feature correlation (↓) on four classification tasks of GPT-3 (Ada, 2.7B), finetuned with and without explanations.

-0.202, respectively). Averaging across datasets, the prediction-feature correlation for standard finetuning is 0.384 while for explanation-based finetuning it is only 0.167 (-0.217). This supports the idea that explanation-based finetuning makes models rely less on spurious cues.

Overall, there is strong evidence to support that including explanations during finetuning can make LLMs more robust to spurious correlations.

6.1 Discussion

Observing the results for CREAK and e-SNLI compared to ComVE and SBIC, it is clear that our approach benefits some tasks more than others.

From Table 2, we see that introducing explanations helps with accuracy the most when the standard-finetuned model has a high prediction-feature correlation. In cases where explanation-based finetuning outperforms standard finetuning on absolute accuracy, the average prediction-feature correlation for the standard finetuning is 0.470; in the opposite case, it is 0.128.

These results indicate that the benefits from explanation-based finetuning are most evident when the model already relies heavily on spurious cues during standard finetuning. When the model does not pick up these cues in the first place, tuning on a set including explanations may have caused the model to underfit the objective of generating the correct binary label, similar to the "no cue" condition. Specifically, each weight update now also has to optimize parts of the network for explanation generation as opposed to only for label generation.

7 Further Analysis

Having shown the effectiveness of our method, we now analyze potential factors that may influence the extent to which it works by answering the following questions:

Do explanations improve the robustness of models of different sizes and families?

We run the same experiments in Section 6 with GPT-3 (Ada), T5, and BART. Table 3 shows the results for GPT-3 (Ada); see Appendix A.2 for T5 and BART results.

We can see that explanations can indeed improve robustness for smaller models as well. In general, though, the improvements are much smaller. For Ada, the absolute accuracy gain from explanation-based finetuning over standard finetuning averaged across all datasets and cues is 1.78, as opposed to 6.85 for Davinci. In terms of Ada's prediction-feature correlation, the average is 0.606 for explanation-based finetuning and 0.728 for standard finetuning. This gap of 0.122 is smaller than that in the case of Davinci, which is 0.217.

Interestingly, when no spurious cue is introduced, adding explanations substantially decreases Ada's accuracy across all datasets by an average of 13.2. On Davinci, this average drop is only 1.75. This suggests that it is more challenging for smaller models to generate good explanations, so the accuracy penalty from explanation-finetuning is reduced as model size increases.

How does the spurious correlation strength affect our method?

As mentioned in Section 4.1, the strength of

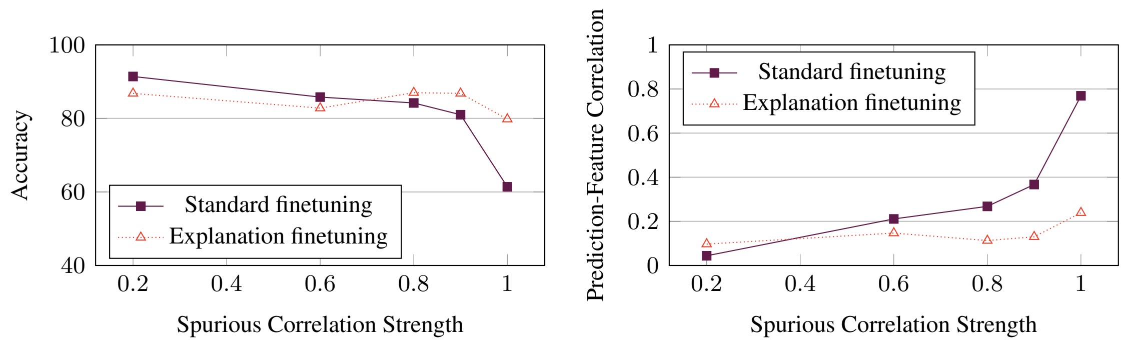

Figure 3: Accuracy (↑) and prediction-feature correlation (↓) of GPT-3 (Davinci) on e-SNLI, as the strength of the "embedding cluster" spurious correlation varies.

the spurious correlation in our skewed training set is maximum. This means that the cue is perfectly correlated with the label (MCC=1.0). Here, we analyze how our method works under different levels of spurious correlation strength in the training set. We select e-SNLI and the embedding cluster cue as a case study. Note that in the main experiments with MCC=1.0, we only sample positive-labeled examples from the pool of examples with the feature present (L_1, f_+) and negative-labeled examples from examples with the feature absent(L_0, f_-). Here, we vary the level of correlation by introducing a certain number of negative-labeled examples exhibiting the feature (L_0, f_+) and positive-labeled examples not exhibiting the feature (L_1, f_-) into the training set.

From Table 2, on e-SNLI, standard-finetuning outperforms explanation-based finetuning on accuracy by 2.4 under the "no cue" condition, where the correlation between the label and the embedding cluster feature is near zero. When the correlation becomes 1.0, this difference is 18.6 in favor of explanation-based finetuning. Between the two extreme cases, we show the results with different levels of spurious correlation strength in Figure 3, in terms of accuracy and prediction-feature correlation. We observe that the explanation-based finetuning starts to perform better than standard finetuning when the correlation between the spurious cue and the target feature is above 0.8, again confirming our finding in Section 6.1.

Does explanation quality affect the effectiveness of our method?

In the in-context learning scenario, Lampinen et al. (2022) show that explanations can improve task performance when used in few-shot prompting. Specifically, they find that high-quality explanations that are manually selected provide substantially more gains than explanations that are not filtered for quality.

To analyze the impact of explanation quality in our setting, we intentionally lower the quality of explanations provided in finetuning by making them irrelevant to the input. We do this via *in-label permutation* on all explanations: for any given instance in the training set, the explanation for its label will be replaced with the explanation from another instance with the *same* label. In other words, the new explanation does not apparently conflict with the label but is irrelevant to the input.

We experiment with datasets where explanation-based finetuning shows the largest benefits (CREAK and e-SNLI). The results are shown in Table 4. Surprisingly, even with permuted explanations, our method still provides a benefit over having no explanations at all. Averaging over all spurious cues and both datasets, the accuracy gain from using permuted explanations compared to having no explanations is 2.85. This is much smaller than the accuracy gain from using the non-permuted explanations (10.25), though.

These results can be compared with the findings from Wang et al. (2022), which shows the central role of explanation relevance in the few-shot setting. To understand why permuted explanations still help in our case, since our data contains spurious cues, we hypothesize that the model could be "distracted" by the explanations even if they are irrelevant and thus "forgets" the spurious cues. We leave it for future work to verify this hypothesis.

Do the explanations have to be human-written?

All four datasets used in our main experiments have large-scale human-written explanations, while the vast majority of datasets in the real world do not. In this analysis, we investigate the possibility of using LM-generated explanations in place of human-written ones, to see if it is possible to

		CREAK			e-SNLI		
		Standard	Explain	Permute	Standard	Explain	Permute
Accuracy (δ_{acc})	No Cue	84.2	85.0	**86.2**	**91.6**	89.2	90.0
	Sentence Length	60.4 (-23.8)	**80.2** **(-4.8)**	67.6 (-18.6)	69.8 (-21.8)	**76.2** **(-13.0)**	72.2 (-17.8)
	Present Tense	74.6 (-9.6)	**80.2** **(-4.8)**	75.4 (-10.8)	85.8 (-5.8)	**88.0** **(-1.2)**	80.2 (-9.8)
	Embedding Cluster	69.2 (-15.0)	**78.6** **(-6.4)**	74.8 (-11.4)	70.6 (-21.0)	**88.6** **(-0.6)**	77.4 (-12.6)
	Average	68.1 (-16.1)	**79.7** **(-5.3)**	72.6 (-13.6)	75.4 (-16.2)	**84.3** **(-4.9)**	76.6 (-13.4)
Prediction-Feature Correlation	Sentence Length	0.847	**0.325**	0.457	0.467	**0.291**	0.382
	Present Tense	0.305	**0.146**	0.319	0.217	**0.143**	0.322
	Embedding Cluster	0.563	**0.288**	-0.427	0.595	**0.141**	-0.303
	Average	0.572	0.253	**0.116**	0.426	0.192	**0.134**

Table 4: Results on CREAK and e-SNLI when explanations are permuted to be completely irrelevant to the input, in comparison with standard finetuning and explanation-based finetuning (with valid explanations).

	Spurious Cues	Standard	Explain	Bootstrap
Accuracy	Sentence Length	60.4	**80.2**	78.6
	Present Tense	74.6	**80.2**	77.2
	Embedding Cluster	69.2	**78.6**	73.0
	Plural Noun	72.2	77.2	**80.8**
	Average	69.1	**79.1**	77.4
Prediction-Feature Correlation	Sentence Length	0.847	0.325	**0.045**
	Present Tense	0.305	0.146	**0.167**
	Embedding Cluster	0.563	**0.288**	0.429
	Plural Noun	0.445	0.170	**0.129**
	Average	0.540	0.232	**0.193**

Table 5: Results on CREAK with bootstrapped explanations generated by the model, in comparison with standard finetuning and explanation-based finetuning (with human-written explanations).

generalize our method to more datasets without explanations.

We perform the experiment on the CREAK dataset as a case study. Specifically, we prompt GPT-3 (Davinci) in a 10-shot setting to generate an explanation for a given input. We do this via a bootstrapping process: (1) we initialize the seed set with 10 training instances, including the label and the human-provided explanation; (2) we sample 10 instances without replacement from the seed set, and prompt the model to generate an explanation for a new instance from the training set; (3) with the generated explanation, we add the new instance to the seed set; (4) we repeat steps (2)-(3) until the entire training set contains explanations. Note that when generating the explanation, we give the model access to the ground-truth label. The temperature is set to 0.9 to facilitate diverse completions.

Results of using these bootstrapped explanations are shown in Table 5. On average, the accuracy gain from finetuning with bootstrapped explanations over no explanations is 8.3. This is slightly lower than the benefit from using human-written explanations (10.0), but still decent. Inspecting the prediction-feature correlation, bootstrapped explanations bring an average correlation drop of 0.347 compared to standard finetuning. This is even greater than the case of using human-written explanations, which is 0.308.

This indicates that explanation-based finetuning can potentially benefit datasets without human-provided explanations, which greatly increases the generalizability and applicability of our approach.

8 Conclusion

We propose explanation-based finetuning, a novel method to reduce model reliance on spurious cues in the training data. In addition to predicting a label, the model is finetuned to generate a free-text explanation in support of its prediction. On a diverse set of classification tasks and spurious features, our method makes the model substantially more robust, as demonstrated by both accuracy and correlation based measures. The efficacy of our method generalizes to different model sizes and families, though larger models tend to benefit more. Moreover, the stronger the spurious correlation occurs in the data, the more helpful our method is. Interestingly, the quality of explanations, in terms of relevance, is not fully necessary as permuted explanations still provide around 25% of the accuracy benefits that the non-permuted explanations provide. What is most notable is that even with model-generated explanations, our method works almost as well as with human-written ones, implying its potential applicability to the vast majority of datasets without explanations available.

9 Limitations

We notice a few key limitations of our approach. The first is, similar to what is found by previous interpretability studies (Camburu et al., 2018, i.a.), incorporating explanations comes with some penalty on in-distribution accuracy, when there is no spurious cue. This penalty decreases as model size increases, though, potentially because it is less challenging for larger models to generate good explanations. The second is that our artificially constructed training set may not be reflective of how strong these spurious cues are in the real world. In our main experiments, we focus on the case where one spurious cue is perfectly correlated to the target label. For further exploration, we can study the alternative setting where there are multiple weak spurious cues instead of a single strong one. Finally, our work here is limited in the scope of experiments. We only experiment with generative LMs and binary classification tasks. Also, because of resource constraints, we only consider four datasets and eight types of spurious cues (including dataset-independent and dataset-specific ones). Additional experiments using a wider variety of spurious cues and datasets would help to shed light on how our method generalizes to other scenarios.

10 Ethics Statement

Potential risks While our work on overcoming spurious cues is related to the idea of debiasing models, it is important to note that our results do not indicate that our method is the best to tackle socially harmful biases against marginalized groups, like gender or racial biases. We have not run any experiments following this direction, and it is important to make this distinction so that the reader does not misunderstand the goal of this paper.

Intended Use Our models and methods shown here are for research purposes only. They should not be deployed in the real world as solutions without further evaluation.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

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. Association for Computational Linguistics.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805. 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.

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. Association for Computational Linguistics.

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

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53. Association for Computational Linguistics.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Haochen Liu, Joseph Thekinen, Sinem Mollaoglu, Da Tang, Ji Yang, Youlong Cheng, Hui Liu, and Jiliang Tang. 2022. Toward annotator group bias in crowdsourcing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1797–1806, Dublin, Ireland. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901. Association for Computational Linguistics.

Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. *arXiv preprint arXiv:2109.01653*.

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. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Frano Rajič, Ivan Stresec, Axel Marmet, and Tim Poštuvan. 2022. Using focal loss to fight shallow heuristics: An empirical analysis of modulated cross-entropy in natural language inference. *arXiv preprint arXiv:2211.13331*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2020. Learning from others' mistakes: Avoiding dataset biases without modeling them. *arXiv preprint arXiv:2012.01300*.

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

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11349–11357.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. There is strength in numbers: Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

A Extended Results

A.1 Results Under "No Cue" Condition

Under the "no cue" condition (i.e., when the training set is unskewed), we report the test accuracy of GPT-3 (Davinci) under finetuning (n=1,000), few-shot (n=10), and zero-shot settings. Results are shown in Table 6. Across the four different datasets, the model finetuned on 1,000 examples achieves much higher accuracies compared to 10-shot or zero-shot prompting.

Comparing standard finetuning and explanation-based finetuning, across all these experiments, we only find an obvious increase (+6.7) on CREAK under the few-shot setting and a slight increase (+0.4) on ComVE under the zero-shot setting. In all other cases, the accuracy either drops or stays the same.

A.2 Results for Other Models

In our main experiments in Section 6 and Section 7, we use OpenAI GPT-3 (Davinci, 175B and Ada, 2.7B), since their relatively large size may allow for the generation of higher-quality experiments, as suggested by (Wei et al., 2022).

We also generalize this approach to other model families including T5-base (220M) and BART-base (110M), which are much smaller generative LMs than GPT-3. Table 7 and Table 8 show the results for these two models respectively. Under the "no cue" condition, their performance is generally much worse than GPT-3 models. The penalty of introducing explanations in finetuning is also more striking, oftentimes resulting in an accuracy around or lower than chance (50.0).

When the training set contains spurious cues, our method still generally works for both T5 and BART on three of the four datasets, as measured by $\delta^{f}_{acc}(M, FT)$ and $corr_f(M^{FT}_f)$. However, the absolute accuracy is almost consistently lower for explanation-based finetuning than for standard finetuning, most likely due to the huge penalty under the "no cue" condition in the first place.

As an exception, on the SBIC dataset, our method does not always work well. For the T5 model, across all spurious features, explanation-based finetuning results in a similar or worse δ_{acc} (the difference is always less than 2.0 percent). It also fails to reduce the prediction-feature correlation for any spurious feature except the "embedding cluster" one, where the correlation only decreases by 0.03. For the BART model, our method

	ComVE		CREAK		e-SNLI		SBIC	
	Standard	Explain	Standard	Explain	Standard	Explain	Standard	Explain
Finetuned (n=1k)	**97.0**	95.5	84.2	**85.0**	**91.6**	89.2	**79.0**	75.0
Fewshot (n=10)	**54.0**	**54.0**	67.5	**74.0**	**59.0**	55.5	**72.0**	66.0
Zero-Shot	47.6	**48.0**	**57.0**	55.5	**51.4**	50.6	56.6	**62.8**

Table 6: Accuracies under the "no cue" condition for all datasets across different finetuning and prompting strategies.

		ComVE		CREAK		e-SNLI		SBIC	
		Standard	Explain	Standard	Explain	Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	No Cue	**76.4**	49.8	**55.2**	41.4	**86.6**	55.6	**69.4**	65.0
	Sentence Length	**53.6** (-22.8)	51.2 **(1.4)**	**52.6** (-2.6)	45.6 **(4.2)**	**64.0** (-22.6)	51.6 **(-4.0)**	**56.0** (-13.4)	53.4 **(-11.6)**
	Present Tense	**61.6** (-14.8)	51.2 **(1.4)**	**50.0** (-5.2)	41.8 **(0.4)**	**79.4** **(-7.2)**	42.6 (-13.0)	**70.6** **(1.2)**	63.6 (-1.4)
	Embedding Cluster	**59.4** (-17.0)	44.6 **(-5.2)**	**49.4** (-5.8)	38.4 **(-3.0)**	**69.8** (-16.8)	42.6 **(-13.0)**	**71.8** **(2.4)**	64.0 (-1.0)
	Plural Noun	**73.8** (-2.6)	53.4 **(3.6)**	**50.8** (-4.4)	40.6 **(-0.8)**	**59.4** (-27.2)	43.8 **(-11.8)**	**69.4** (0.0)	66.4 **(1.4)**
	Average	**62.1** (-14.3)	50.1 **(0.3)**	**50.7** (-4.5)	41.6 **(0.2)**	**68.2** (-18.5)	45.2 **(-10.5)**	**67.0** **(-2.5)**	61.9 (-3.2)
Prediction-Feature Correlation	Sentence Length	0.641	**0.402**	0.699	**0.115**	0.524	**0.384**	**0.222**	0.376
	Present Tense	0.653	**0.166**	0.575	**0.513**	0.281	**0.231**	**0.217**	0.319
	Embedding Cluster	0.645	**0.463**	0.694	**0.456**	0.494	**0.169**	0.504	**0.473**
	Plural Noun	0.343	**0.176**	0.481	**0.269**	0.722	**0.207**	**0.107**	0.205
	Average	0.571	**0.302**	0.612	**0.338**	0.505	**0.248**	**0.263**	0.343

Table 7: Accuracy (↑), accuracy drop (↑), and prediction-feature correlation (↓) on four classification tasks of T5-base, finetuned with and without explanations.

does make it more robust to the "embedding cluster" and the "plural noun" cues but no other cues, as reflected by both the accuracy drop and the prediction-feature correlation. We hypothesize that this is because of the model does not rely heavily on the cues in the first place, as shown by the lower prediction-feature correlations in the case of standard finetuning. This reconfirms our observation from Section 6.1. Generally, compared to GPT-3, our method still works on most of the datasets for T5 and BART, but with smaller benefits. This is most likely because explanation generation is in itself a challenging task for smaller models, thus resulting in a larger penalty on accuracy in the "no cue" condition.

B Extended Analysis

B.1 Does knowledge of the cue improve model robustness via few-shot prompting?

In our main experiments, we only consider datasets that come with human-annotated explanations for all training instances. However, this is untrue for the vast majority of datasets in the real world. Here, we want to explore if it is possible to overcome the cue *without* large-scale human-written explanations available. Specifically, given only a few examples of human-written explanations, can we still make the model more robust, if we have knowledge about what the spurious feature is?

Specifically, we take standard-finetuned models trained on the skewed training sets. Then, we use 10 training examples to construct the few-shot prompt. In the standard prompting setting, we only include the input and the label; for explanation-based prompting, we additionally include a free-text explanation before the label. For both settings, the set of few-shot examples is randomly shuffled and unskewed (i.e., they do not exhibit the spurious correlation).

We experiment with e-SNLI in this analysis. The results, as shown in Table 9, indicate that for syntactic spurious cues, standard prompting significantly helps the standard model become more robust to

		ComVE		CREAK		e-SNLI		SBIC	
		Standard	Explain	Standard	Explain	Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	No Cue	**53.2**	48.0	**59.0**	46.0	**85.6**	46.2	**76.8**	51.0
	Sentence Length	42.8 (-10.4)	**43.6** **(-4.4)**	**54.6** (-4.4)	48.4 **(2.4)**	**54.4** (-31.2)	43.4 **(-2.8)**	**50.8** (-26.0)	50.2 **(-0.8)**
	Present Tense	**55.2** (2.0)	54.0 **(6.0)**	**53.2** (-5.8)	48.0 **(2.0)**	**58.8** (-26.8)	44.0 **(-2.2)**	**70.8** (-6.0)	63.0 **(12.0)**
	Embedding Cluster	**48.0** (-5.2)	47.2 **(-0.8)**	**49.6** (-9.4)	44.0 **(-2.0)**	**54.0** (-31.6)	40.6 **(-5.6)**	**68.6** (-8.2)	60.0 **(9.0)**
	Plural Noun	**54.0** (0.8)	51.2 **(3.2)**	**53.2** (-5.8)	46.8 **(0.8)**	**52.8** (-32.8)	48.4 **(2.2)**	**65.2** (-11.6)	53.8 **(2.8)**
	Average	**50.0** (-3.2)	49.0 **(1.0)**	**52.7** (-6.4)	46.8 **(0.8)**	**55.0** (-30.6)	44.1 **(-2.1)**	**63.9** (-13.0)	56.8 **(5.8)**
Prediction-Feature Correlation	Sentence Length	0.667	**0.638**	0.762	**0.629**	**0.724**	0.745	**0.288**	0.706
	Present Tense	0.881	**0.744**	0.603	**0.454**	0.702	**0.159**	**0.241**	0.314
	Embedding Cluster	0.817	**0.792**	0.801	**0.700**	0.854	**0.301**	0.555	**0.395**
	Plural Noun	0.823	**0.230**	0.607	**0.491**	0.884	**0.439**	0.287	**0.210**
	Average	0.797	**0.601**	0.693	**0.569**	0.791	**0.411**	**0.343**	0.406

Table 8: Accuracy (↑), accuracy drop (↑), and prediction-feature correlation (↓) on four classification tasks of BART-base, finetuned with and without explanations.

them. The correlations between the model predictions and the spurious cue drop by 0.297 to 0.359 for the three syntactic spurious cues. However, there is no evidence that few-shot prompting benefits when the "embedding cluster" cue is introduced. Although adding explanations is shown to be effective in finetuning, it does not help as much in few-shot prompting, in terms of either accuracy or prediction-feature correlation.

B.2 Increasing the number of finetuning examples from 1k to 4k

In this analysis, we examine the effect of increasing the number of training examples for finetuning from 1k to 4k. This is to investigate the hypothesis that increasing the number of training examples will make it easier for models to learn, and subsequently overfit on the spurious cue.

e-SNLI Experiments. We repeat the experiments used to create Figure 3 with the modification that instead of being trained on 1k examples, models are trained on 4k examples. These results are shown in Table 10. In the table, we find that the accuracies of both the standard and finetuned models improve when we increase the number of training examples. The average standard finetuning model increases by 2.3 while for the explanation-based finetuned models this increase is 5.2. Correspondingly, the average accuracy gap also increases between the standard and explanation-based models from 4.52 in the n=1k to 6.70 (+2.18).

Looking at the prediction-feature correlation, we note that the average correlation does not change substantially for both the standard finetuning and explanation finetuning after increasing the number of training examples to 4k.

Overall, these results provide evidence that having an increased number of examples tends to benefit both standard and explanation based finetunes with explanation-based finetunes being able to benefit more. However, in the case that the training set correlation between the target label and the spurious cue is 1.0, we note that the performance for the standard finetuning drops substantially.

ComVE and SBIC Experiments. Furthering the results from the previous analysis, we investigate the effect of increasing the number of finetuning examples in the cases where we found the effect of explanation-based finetuning to be the weakest in Table 2. Specifically, we investigate SBIC and ComVE under the present tense and sentence length spurious cues by rerunning the experiments under this setting with the modification of increasing the training set size from 1k to 4k. These results are shown in Table 11.

These results provide strong confirmation that increasing the number of examples when the spurious cue is perfectly correlated with the label substantially degrades model performance. Under the setting where we only have 1k training examples, the average accuracy difference between standard and explanation-based finetuning across both cues and datasets is 1.0 in favor of standard finetuning. This difference when we have 4k training examples

		Standard-finetuned model	Standard prompting on standard-finetuned model	Explanation-based prompting on standard-finetuned model
Accuracy (δ_{acc})	No cue	88.0	N/A	N/A
	Sentence Length	69.8 (-18.2)	**78.4 (-9.6)**	72.4 (-15.6)
	Present Tense	76.0 (-12.0)	**86.0 (-2.0)**	83.6 (-4.4)
	Embedding Cluster	70.6 (-17.4)	**71.2 (-16.8)**	62.8 (-25.2)
	Plural Noun	69.0 (-19.0)	**83.6 (-4.4)**	78.6 (-9.4)
	Average	71.4 (-16.7)	**79.8 (-8.2)**	74.4 (-13.7)
Prediction-Feature Correlation	Sentence Length	0.467	**0.109**	0.148
	Present Tense	0.336	**0.039**	0.085
	Embedding Cluster	0.595	**0.532**	0.691
	Plural Noun	0.578	**0.219**	0.304
	Average	0.494	**0.225**	0.307

Table 9: Standard few-shot prompting vs. explanation-based few-shot prompting on standard-finetuned model with e-SNLI dataset. The accuracy difference δ_{acc} for the last two columns are based on the standard-finetuned model under the "no cue" setting.

		1k Examples		4k Examples	
		Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	0.2	**91.4**	86.8	**94.4**	89.8
	0.6	**85.8**	82.8	**90.8**	90.6
	0.8	84.2	**87.0**	88.4	**89.8**
	0.9	81.0	**86.8**	83.2	**91.0**
	1.0	61.4	**79.8**	58.8	**87.6**
	Average	80.8	**84.6**	83.1	**89.8**
Prediction-Feature Correlation	0.2	**0.044**	0.097	**0.024**	0.094
	0.6	0.211	**0.147**	0.160	**0.117**
	0.8	0.268	**0.113**	0.231	**0.158**
	0.9	0.367	**0.130**	0.336	**0.141**
	1.0	0.769	**0.239**	0.84	**0.233**
	Average	0.332	**0.145**	0.318	**0.149**

Table 10: Accuracy (↑) and prediction-feature correlation (↓) of GPT-3 (Davinci) on e-SNLI, as the strength of the "embedding cluster" spurious correlation and the number of training examples varies.

			ComVE		SBIC	
			Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	Present Tense	n=1k	**93.6**	89.4	**78.6**	77.4
		n=4k	81.6	**94.8**	65.4	**77.0**
	Sentence Length	n=1k	**91.4**	89.4	50.4	**53.4**
		n=4k	83.2	**89.0**	50.4	**53.4**
Prediction-Feature Correlation	Present Tense	n=1k	0.074	**0.035**	0.241	**0.166**
		n=4k	0.316	**0.021**	0.387	**-0.001**
	Sentence Length	n=1k	0.134	**0.108**	0.732	**0.166**
		n=4k	0.245	**0.109**	0.770	**0.670**

Table 11: Standard finetuning vs. explanation-based finetuning on selected settings after increasing number of examples.

is 8.4 in favor of explanation-based finetuning. It is worth noting that in three out of the four settings in this experiment (everything except length bias for SBIC), in the n=1k setting, standard finetuning does not provide a benefit. However, if we increase n to 4k, that increases the model's susceptibility to the cue enough that explanation-based finetuning outperforms standard finetuning, a reversal of the original results.

B.3 Dataset-Specific Spurious Cues

In addition to the four common spurious cues in the main text, we also construct dataset-specific spurious correlations to simulate realistic cues that can naturally appear in each dataset:

Higher Perplexity (CREAK). Using GPT-2 to measure perplexity, we filter the data into a set with above-median perplexity and a set with below-median perplexity. This feature is considered to be present if the perplexity of the sentence is higher than the median perplexity and is positively labeled.

Gender Female (e-SNLI). If the premise contains female-related pronouns (woman, women, girl, lady, etc.), we consider the "gender female" spurious cue to be present. The aforementioned words frequently appear in the e-SNLI dataset when the sentence is relevant to females.

Username Mentions (SBIC). If the social media post contains an "@" sign, meaning the author might be tagging or directly replying to other users on social media, we consider the spurious cue to be present. This feature is supposed to have no causal relationship with whether a post is offensive.

		ComVE		CREAK		e-SNLI		SBIC	
		Standard	Explain	Standard	Explain	Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	No Cue	**97.0**	95.6	84.2	**85.0**	**91.6**	89.2	**79.0**	75.0
	Domain Specific	**93.6** **(-3.4)**	90.4 (-5.3)	**80.5** **(-3.7)**	79.0 (-6.0)	55.8 (-35.8)	**86.6** **(-2.6)**	**42.6** **(-36.4)**	38.3 (-36.7)
Prediction-Feature Correlation	Domain Specific	**0.055**	0.097	0.112	**-0.026**	0.684	**0.080**	0.991	**0.915**

Table 12: Accuracy (↑), accuracy drop (↑), and prediction-feature correlation (↓) on four classification tasks of GPT-3 (Davinci, 175B), finetuned with and without explanations. The skewed training sets contain domain-specific cues.

POS-tag of Swapped Word (ComVE). The ComVE dataset requires us to compare two sentences and output which sentence makes more sense, the two sentences have high lexical overlaps. We consider the part of speech (POS) of the first word which is different between the two sentences and say that the POS tag of swapped word spurious cue is present if this word is a noun.

Table 12 shows the performance of GPT-3 (Davinci). When adding "gender female" spurious cues to the e-SNLI dataset, we find strong evidence that explanations make the model less susceptible to the spurious cue. In standard finetuning, the prediction-feature correlation is 0.684 and the accuracy is 55.8, suggesting the model relies heavily on the spurious pattern. Meanwhile, for the model finetuned with explanations, this correlation drops to 0.080, and the accuracy increases to 86.6. The results for dataset-specific cues of the ComVE and CREAK datasets are consistent with our finding that our approach is most effective when the spurious cues highly impact the model performance. On the SBIC dataset, explanation-based finetuning only decreases the prediction-feature correlation by 0.076. This could be due to the fact that the "username mention" cue is the most shallow one among all domain-specific cues, since the model only needs to detect one token ("@"), which makes it surprisingly easy for it to pick up the cue.

C Implementation Details

C.1 Spurious Cue Implementation

The implementation of the "present tense" and "plural noun" spurious cues described in Section 5.2 and the "POS-tag of swapped word" cue in the Section B.3 involve tokenizing and performing POS tagging on the inputs. The tokenizer and POS-tagger we use are implemented by (Bird et al., 2009) in the NLTK toolkit [7].

For the "higher perplexity" spurious cue for the CREAK dataset, we compute the GPT-2 perplexity of the input text using the metric module implemented in the Huggingface Evaluate package [8]. Its license is Apache License 2.0.

C.2 Models and Hyperparameters

All our code are attached as the supplemental materials.

OpenAI Models We finetuned GPT-3 (Brown et al., 2020) from OpenAI's standard API[9] in different sizes (`Davinci` and `Ada`). Its license is MIT license. The GPT-3 models are finetuned for four epochs (default setting on the OpenAI API), and the other hyperparameters (e.g. learning rates) are the default values. with the exception of the models trained with 4k examples which were only trained for one epoch with an increased learning rate (0.2) to reduce costs.

Huggingface Models T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) are implemented with HuggingFace Transformers[10]. The pretrained model checkpoints we use are the `t5-base` (220M parameters) and `facebook/bart-base` (110M parameters). Their licenses are Apache License 2.0. We use the `conditional generation` classes for T5 [11] and BART [12] from Huggingface to finetune the pretrained models. To remain consistent with the finetuning of OpenAI models, the T5 and BART models are finetuned with 1,000 training examples

[7] https://www.nltk.org/
[8] https://github.com/huggingface/evaluate
[9] https://beta.openai.com/docs/api-reference
[10] https://github.com/huggingface/transformers
[11] https://huggingface.co/docs/transformers/model_doc/t5#transformers.T5ForConditionalGeneration
[12] https://huggingface.co/docs/transformers/model_doc/bart#transformers.BartForConditionalGeneration

and run for 4 training epochs. The batch size is set to 8 and the learning rate is set to 2e-5 with the max sequence length being 128. Our finetuning experiments are run on a Kepler K80 GPU. Each finetuning takes 5 to 10 minutes depending on the task.

C.3 Computational Resources

All experiments performed using GPT-3 including all finetuning were performed using the OpenAI public API. We note that every finetuning experiment on each cue and dataset in this paper costs around $10 to perform. Across all our datasets, creating a finetuned model involving 1k samples cost around $5 when tuned without explanations and $7 with explanations. Performing evaluation with these finetuned models then cost around a dollar when evaluating on 500 samples.

All other experiments involving heavy computational resources such as finetuning T5 and BART were performed on Google Colaboratory with GPU-accelerated notebooks available on the pro subscription.

D Datasets Details

D.1 Dataset URLs and Licenses

Listed below are all the details and licenses (where available) for the datasets used in this paper. All datasets used were research datasets and used for their intended purposes. None of the data used in this paper contains any sensitive information. A disclaimer has been added at the start of this paper for offensive content given that the SBIC dataset contains examples of hate speech.

CREAK (Onoe et al., 2021) : https://github.com/yasumasaonoe/creak

e-SNLI (Camburu et al., 2018) : https://github.com/OanaMariaCamburu/e-SNLI, license: MIT License, https://github.com/OanaMariaCamburu/e-SNLI/blob/master/LICENSE

ComVE (Wang et al., 2019) : https://github.com/wangcunxiang/SemEval2020-Task4-Commonsense-Validation\-and-Explanation, license: CC BY 4.0

SBIC (Sap et al., 2020) : https://maartensap.com/social-bias-frames/SBIC.v2.tgz, license: CC BY 4.0

	ComVE	CREAK	e-SNLI	SBIC
Sentence Length	0.018	0.056	-0.114	-0.226
Present Tense	-0.022	-0.010	-0.004	0.135
Embedding Cluster	0.000	-0.008	0.062	0.378
Plural Noun	-0.062	0.006	0.007	0.112
Dataset-specific	-0.051	-0.004	-0.059	-0.068

Table 13: Label-feature correlation in the unskewed training set D_{train} without intentionally introduced spurious cues.

D.2 Label-Feature Correlation in Unskewed Training Sets

The correlation between the ground-truth label and the spurious cues on the randomly selected 1,000 training sets is shown in Table 13. There are no artificially introduced spurious correlations in this training set. According to the correlations in the table, we claim that the “no cue” training set is unskewed, except for the “embedding cluster” on the SBIC dataset where this correlation is 0.378, implying that the embedding vectors for the offensive social media posts are clustered together.

		ComVE		CREAK		e-SNLI		SBIC	
		Standard	Explain	Standard	Explain	Standard	Explain	Standard	Explain
Accuracy (δ_{acc})	No Cue	**87.4**	74.0	**76.8**	68.6	**90.4**	86.0	78.0	**78.6**
	Sentence Length	50.4 (-37.0)	**59.0** **(-15.0)**	52.8 (-24.0)	**59.4** **(-9.2)**	**62.2** (-28.2)	60.6 **(-25.4)**	51.2 (-26.8)	**52.0** **(-26.6)**
	Present Tense	54.2 (-33.2)	**69.6** **(-4.4)**	55.8 (-21.0)	**61.6** **(-7.0)**	75.0 (-15.4)	**76.4** **(-9.6)**	73.6 (-4.4)	**75.6** **(-3.0)**
	Embedding Cluster	50.8 (-36.6)	**55.6** **(-18.4)**	51.8 (-25.0)	**55.4** **(-13.2)**	63.2 (-27.2)	**69.0** **(-17.0)**	54.8 (-23.2)	**56.6** **(-22.0)**
	Plural Noun	52.8 (-34.6)	**64.8** **(-9.2)**	54.4 (-22.4)	**62.6** **(-6.0)**	59.8 (-30.6)	**63.6** **(-22.4)**	75.8 (-2.2)	**78.2** **(-0.4)**
	Average	52.1 (-35.4)	**62.3** **(-11.8)**	53.7 (-23.1)	**59.8** **(-8.8)**	65.1 (-25.4)	**67.4** **(-18.6)**	63.9 (-14.2)	**65.6** **(-13.0)**
Correlation between Model's Prediction and Spurious Feature	Sentence Length	0.821	**0.524**	0.894	**0.659**	0.633	**0.582**	0.753	**0.735**
	Present Tense	0.791	**0.528**	0.704	**0.465**	0.439	**0.341**	0.417	**0.269**
	Embedding Cluster	0.815	**0.675**	0.735	**0.665**	0.761	**0.484**	0.570	**0.551**
	Plural Noun	0.838	**0.494**	0.714	**0.373**	0.721	**0.579**	0.220	**0.191**
	Average	0.816	**0.555**	0.762	**0.541**	0.639	**0.496**	0.490	**0.437**

Table 14: Accuracy (↑), accuracy drop (↑), and prediction-feature correlation (↓) on four classification tasks of GPT-3 (Babbage), finetuned with and without explanations.