Adversarial Robustness in LLMs: Evaluating Two-Step Fine-Tuning and Spurious Correlation Extraction Methods

Anonymous ACL submission

Abstract

Recently, Large Language Models (LLMs) have achieved remarkable progress in text classification. However, they remain prone to spurious correlations or shortcuts—unintended associations between training data and task labels—that can degrade generalization and adversarial robustness. Most existing approaches identify a limited set of task-specific shortcuts using human priors or data augmentation, which require extensive expertise and manual effort. In this study, we investigate a two-step fine-tuning approach to mitigate such spurious correlations and improve model reliability. Our method involves an initial fine-tuning phase on the full dataset, followed by a second fine-tuning phase on a subset identified as spurious using an extraction technique among important scores or LID scores. We evaluate our approach on four classification datasets (MNLI, HANS, FEVER, and FEVER-Symmetric) and find that this informed fine-tuning strategy enhances robustness. Our methods following attention scores prove to be the effective in identifying spurious correlations. Notably, with a BERT-base model, we achieve an accuracy of 73.6% on HANS, compared to 62.9% with standard fine-tuning.[1]

1 Introduction

LLMs, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have achieved strong performance in natural language inference (NLI) tasks but remain vulnerable to spurious correlations (Ye et al., 2024; Geirhos et al., 2020). These correlations also known as shortcuts arise when features in the input predict the label in training data but fail to generalize under distribution shifts (Makar et al., 2022; Honda et al., 2024). A spurious correlation refers to an apparent relationship between two variables that is not truly causal; rather, any observed dependencies arise either by chance or due to an unobserved confounding factor that influences both variables. For example, the term "Spielberg" may correlate with positive sentiment due to frequent positive reviews of his movies, even though the term itself does not convey sentiment (Wang and Culotta, 2020).

Existing mitigation methods rely heavily on human annotations or data augmentation, which are often task-specific and resource-intensive. To address this, we propose a two-step fine-tuning approach: (1) an initial fine-tuning on the full dataset, followed by (2) a targeted fine-tuning phase on data identified as spurious using predefined extraction methods.

We compare our results evaluate three extraction techniques—forgettable examples, attention scores, and LID scores—on four classification datasets (MNLI, HANS, FEVER, and FEVER-Symmetric). We build upon the work of Yaghoobzadeh et al. (2021), who use forgettable examples as an extraction method. Our approach outperforms their results on the HANS dataset using a BERT model, achieving 73.6% compared to their 68.9% for the forget event.

2 State of the Art

To increase robustness to spurious correlations, several methods have been explored. First, it is crucial to define an effective identification method. Once identified, spurious correlations can either be mitigated or leveraged for better understanding. Strategies include data augmentation, adversarial learning, transfer learning, counterfactual generation, or the use of Pre-trained Language Models (PLMs) with refined fine-tuning techniques.

Wang and Culotta (2020) introduces a supervised method leveraging treatment effect estimators to detect spurious correlations at the word level. However, this approach requires human annotation,

[1] The code is avaible at https://anonymous.4open.science/r/robustness_to_spurious_correlation-00FE/README.md

which limits scalability. (Wang et al., 2022) addresses this with an automated framework using attention scores and integrated gradients to differentiate genuine from spurious correlations via cross-dataset analysis.

An other method consist of using **Pre-trained Language Models (PLMs)**, improving robustness by generalizing from minority examples (Tu et al., 2020). Multi-task learning further enhances performance by leveraging auxiliary tasks (Yu et al., 2024). Some works propose alternatives like adding richer contexts or reconsidering **fine-tuning** strategies (Schwartz and Stanovsky, 2022).

Transfer learning enhances generalization, and **adversarial learning** exposes models to challenging inputs, improving robustness (Nie et al., 2020). Datasets like HANS test models reliance on heuristics (McCoy et al., 2019).

Our study is build on the work dealing with **Forgettable examples**, introduced by Yaghoobzadeh et al. (2021), offers a systematic, data-driven approach to mitigating shortcut learning. This approach builds on the concept of example forgetting (Toneva et al., 2019), where instances that are repeatedly misclassified or learned and later forgotten during training highlight hard-to-learn patterns and minority cases. This study introduce a new approach to robustify models by fine-tuning the model twice, first on the full training data and second on the minorities only. Unfortunaltely, this method does not account for all spurious correlations. Indeed, a spurious correlation may consistently be misclassified without its classification ever changing. This reveals the need for a more effective method.

3 Methods

Our approach is a two-stage process outlined in Figure 1.

- Fine-tune BERT model on the entire dataset (e.g., MNLI) for four epochs to establish a baseline. And concurrently, identify spurious correlations with a pre set method (important score, LID score, forget events) across training epochs.

- Fine-tune the trained BERT model exclusively on the extracted examples to mitigate spurious correlations and enhance robustness.

For the extraction of spurious correlations, we consider a balanced approach by class. This means that we equalize the spurious correlations across all classes, thus avoiding dis-proportionality or overfitting of any single class.

To identify spurious correlation features, this study consider two methods :

We propose to identify spurious correlation through **important score**. Defining importance score for an example in a simplified formula:

$$"Importance = (1 - \text{Accuracy}) + \text{Loss}" \quad (1)$$

The proposed importance score is based on two key factors: the misclassification rate "$1 - accuracy$" and the loss value. A high misclassification rate indicates that the model struggles to learn the underlying features of an example, while a high loss suggests uncertainty in the prediction, even when the example is correctly classified. This uncertainty may signal an outlier or an underrepresented instance in the training set. By combining these factors, the importance score highlights both misclassified samples and uncertain predictions, making it a robust measure of data difficulty. It effectively captures dataset complexity by detecting examples that require better feature representation, identifying overconfident errors where incorrect predictions leads to low loss, and distinguishing instances that are susceptible to spurious correlations.

Figure 2 in the Appendix A illustrates the number of examples based on their importance score. Examples with a score close to zero are not shown, as they dominate the distribution. The decaying trend suggests that most examples are "easy" for the model, while a smaller subset poses greater difficulty.

Important features through **Local Intrinsic Dimensionality (LID) score** (Savić et al., 2023) : In this section, we primarily address the fact that spurious correlations often arise from irrelevant or noisy features.

LID score quantifies the perceived dimensionality of neighboring points in a space, essentially measuring the local complexity around a point. Specifically, it assesses how many dimensions are required to explain the variability of data in the vicinity of a given point.

Spurious correlations introduce additional variability into the data, which can increase the local dispersion of points in the representation space. Consequently, the LID may be higher in such cases.

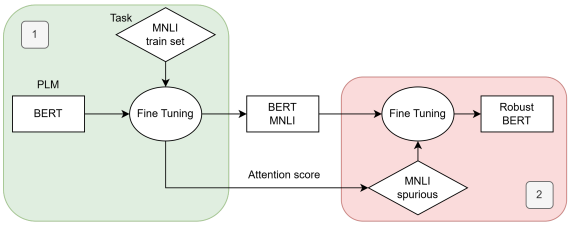

Figure 1: Two step fine tuning method on MNLI task with identification of spurious correlation through important score

However, certain limitations become apparent. In some cases, spurious correlations can lead to a low LID. For instance, if the model relies on a simple cue or an isolated feature that captures low local dimensionality (e.g., recognizing a camel based on the presence of sand in the background). There is no causal relationship here: a high or low LID alone does not prove that a correlation is spurious; it merely highlights areas where local relationships are either simple or complex.

Given our coarse assumption on spurious correlation, we propose to combine these method, for a better chance to fit the real spurious correlation.

4 Experiments and Results

4.1 Model

This study focuses on BERT model (Devlin et al., 2019). We choose BERT for its major use in previous work (Yaghoobzadeh et al., 2021), (Clark et al., 2019) (Utama et al., 2020) (He et al., 2023). BERT is a Transformer, which follow the encoder-decoder framework using stacked multi-head self-attention and fully connected layers for both the encoder and decoder. We used the pre-trained *googlebert/bert-base-cased*[2] model with 12 layers and 12 attention heads.

4.2 Tasks and Datasets

This study explore two natural language understanding (NLU) tasks: natural language inference (NLI) and fact verification. For NLI, we consider the **MNLI** dataset (Williams et al., 2018), which consists of sentence pairs labeled as entailment, contradiction, or neutral, indicating their semantic relationship.

We evaluate models on both, In-distribution (ID) evaluation with the MNLI-dev set and Out-of-distribution (OOD) evaluation with **HANS** dataset (Heuristic Analysis for NLI Systems) (McCoy et al., 2019), to test models robustness and generalization. The HANS dataset contains sentence pairs with high lexical overlap (e.g., "The president advised the doctor" vs. "The doctor advised the president"). This dataset highlights how NLI models tend to over-rely on lexical overlap as a shortcut, incorrectly predicting entailment when two sentences share many words.

For fact verification, we use the **FEVER** dataset (Thorne et al., 2018), where the task is to assess the validity of a claim based on provided evidence. Claims are labeled as Supported, Refuted, or Not Enough Information. To evaluate model robustness, we also test on the **FEVER symmetric** dataset (Schuster et al., 2019), which was designed to mitigate dataset biases by ensuring that models cannot rely on surface cues alone.

In both tasks, models are evaluated using accuracy as the primary metric.

4.3 Results

This part is divided between NLI and fact verification task. We demonstrate that a two step fine tuning and using important score lead to satisfying results. First, Table 1 shows results for **NLI** task. The notation "$\mathcal{FT} + \mathcal{FT}_{important}$" refers to an initial fine-tuning of BERT on MNLI, followed by a second fine-tuning on the examples extracted with important score. The results are satisfactory, using important and LID score and the two-stage fine-tuning process effectively enhances the robustness of the BERT model. Although there is a slight decrease in accuracy on MNLI, this difference is negligible compared to the improvement observed on the HANS dataset. We have set a threshold of 10% of examples extracted on the full dataset. Table 1 also present the results for the combination of important score and LID score. To do so, we consider the union between important examples and forgetting events, thus removing the duplicate examples. Interestingly, this approach does not significantly improve the accuracy on HANS, as the best result remains achieved with important samples. Note that, the accuracy on HANS reaches 72.34% after the first epoch, which is relatively high for the first epoch compared to the previous model. For a comparison, HANS accuracy for the epoch 1 and 2 on $BERT + \mathcal{F}_{important}$ are respectively 69.5 and 70.1. This result is merely due to the number of example that is almost double for the combination.

[2] https://huggingface.co/google-bert/bert-base-uncased

Model	MNLI	HANS	Avg.
$\mathcal{FT}$	84.2	62.9	73.5
$\mathcal{FT} + \mathcal{FT}_{important}$	82.7	73.6	**78.1**
$\mathcal{FT} + \mathcal{FT}_{LID}$	82.4	73.2	77.8
$\mathcal{FT}+\mathcal{FT}_{important \cup LID}$	82.5	73.4	77.9

Table 1: Models accuracy on MNLI and HANS datasets

In Table 2, we report the results of our method applied to the FEVER and symmetric evaluation sets. Our approach again works well for both important and LID score.

Overall, our study achieves highly promising results, particularly on the HANS and FEVER datasets, as well as on the symmetry-based evaluation benchmarks (Symm 1 and Symm 2). Our method demonstrates significant improvements in robustness over prior approaches, highlighting the effectiveness of our proposed strategies. Notably, it achieves results that surpass previous studies (see Table 4 in Appendix A). Moreover, the proposed method is both easy to implement and computationally efficient.

4.4 Analyses

Figure 3 presents a comparison of three extraction methods: forget events (Toneva et al., 2019), importance scores, and LID scores. We observe that training on forget examples leads to a performance decline after the first epoch, suggesting potential overfitting. In contrast, examples extracted using attention scores or LID yield promising results, improving performance by approximately 7% over the baseline model.

We also report the number of extracted examples and their distribution across labels depending on the extraction method used. We observe that, for the MNLI dataset, the examples classified as spurious mainly come from the neutral class, regardless of the extraction criterion.

The predominance of the neutral class among the extracted examples may suggest that these examples are more likely to be affected by spurious correlations. It seems that neutral examples are more ambiguous. This analysis reveals a notable trend: the neutral class is overrepresented among examples affected by spurious correlations. This suggests a potential bias in how these examples are learned and utilized by the model.

5 Conclusion

We proposed a two-step fine-tuning approach to mitigate spurious correlations by leveraging LID and attention scores for targeted example extraction. By identifying and refining a subset of challenging examples, our method enables models to focus on more reliable patterns during training. Evaluating our approach on BERT across NLI and fact verification tasks, we observed consistent gains in robustness, demonstrating its effectiveness in improving model generalization to challenging cases.

Carbon Impact Statement. This work contributed 383 g of carbon equivalent across all experiments and used 2.155 kWh of electricity. Result delivered by CodeCarbon[3].

[3] https://codecarbon.io/

Model	FEVER	Sym-1	Sym-2	Avg.
$\mathcal{FT}$	85.6	56.6	61.0	73.5
$\mathcal{FT} + \mathcal{FT}_{important}$	85.8	63.5	67.2	72.2
$\mathcal{FT} + \mathcal{FT}_{LID}$	87.3	62.9	67.6	72.6

Table 2: Models accuracy on FEVER and FEVER symmetric 1 and 2

Label	Important	LID
neutral	10 766	10 544
contradiction	7 801	7 714
entailment	7 797	7 688

Table 3: Number of spurious examples extracted per method and their label distribution for MNLI

References

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don‘t Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673. ArXiv:2004.07780 [cs].

Zirui He, Huiqi Deng, Haiyan Zhao, Ninghao Liu, and Mengnan Du. 2023. Mitigating Shortcuts in Language Models with Soft Label Encoding. *arXiv preprint*. ArXiv:2309.09380 [cs].

Ukyo Honda, Tatsushi Oka, Peinan Zhang, and Masato Mita. 2024. Not Eliminate but Aggregate: Post-Hoc Control over Mixture-of-Experts to Address Shortcut Shifts in Natural Language Understanding. *Transactions of the Association for Computational Linguistics*, 12:1268–1289. ArXiv:2406.12060 [cs].

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-End Bias Mitigation by Modelling Biases in Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

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. *arXiv preprint*. ArXiv:1907.11692 [cs].

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D’Amour. 2022. Causally motivated shortcut removal using auxiliary labels. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR. ISSN: 2640-3498.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv preprint*. ArXiv:1902.01007 [cs].

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. *arXiv preprint*. ArXiv:1910.14599 [cs].

Miloš Savić, Vladimir Kurbalija, and Miloš Radovanović. 2023. Local Intrinsic Dimensionality Measures for Graphs, with Applications to Graph Embeddings. *arXiv preprint*. ArXiv:2208.11986 [cs].

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Roy Schwartz and Gabriel Stanovsky. 2022. On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations. *arXiv preprint*. ArXiv:2204.12708 [cs].

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An Empirical Study of Example Forgetting during Deep Neural Network Learning. *arXiv preprint*. ArXiv:1812.05159 [cs].

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models.

Transactions of the Association for Computational Linguistics, 8:621–633. Place: Cambridge, MA Publisher: MIT Press.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models. *arXiv preprint*. ArXiv:2110.07736 [cs].

Zhao Wang and Aron Culotta. 2020. Identifying Spurious Correlations for Robust Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *arXiv preprint*. ArXiv:1704.05426 [cs].

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing Robustness to Spurious Correlations using Forgettable Examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024. Spurious Correlations in Machine Learning: A Survey. *arXiv preprint*. ArXiv:2402.12715 [cs].

Jialin Yu, Yuxiang Zhou, Yulan He, Nevin L. Zhang, and Ricardo Silva. 2024. Fine-Tuning Pre-trained Language Models for Robust Causal Representation Learning. *arXiv preprint*. ArXiv:2410.14375 [cs].

A Appendix

In this appendix, we present additional figures and tables that support our main findings.

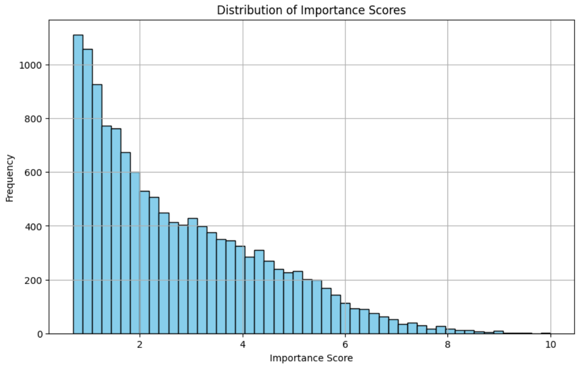

Figure 2: Distribution of important scores for token selection.

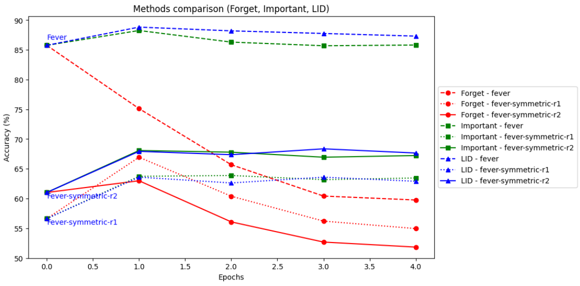

Figure 3: Accuracy comparison of models fine-tuned with different extraction methods across FEVER datasets.

Model	**MNLI**	**HANS**	**Avg.**	**FEVER**	**Sym-1**	**Sym-2**	**Avg.**
BERT (original)	84.2	69.5	73.5	85.6	56.6	61.0	67.7
Post-HOC, uniform (Honda et al., 2024)	83.0	63.6	73.3	87.6	62.2	68.2	72.7
Post-HOC, argmin (Honda et al., 2024)	81.0	67.2	74.1	85.3	61.8	67.4	71.5
Reweighting (Schuster et al., 2019)	81.4	68.6	75.0	84.6	61.7	64.9	70.4
PoE (Clark et al., 2019; Karimi Mahabadi et al., 2020)	84.2	64.6	74.4	82.3	62.0	64.3	69.5
Reg-conf (Utama et al., 2020)	**84.3**	69.1	76.7	86.4	60.5	66.2	69.2
SoftLE (He et al., 2023)	81.2	68.1	74.6	87.5	60.3	66.9	71.6
$\mathcal{FT} + \mathcal{FT}_{forget}$ (Yaghoobzadeh et al., 2021)	83.0	68.9	75.9	87.1	61.0	67.0	71.7
$\mathcal{FT} + \mathcal{FT}_{important}$	82.7	**73.6**	**78.1**	85.8	**63.5**	67.2	72.2
$\mathcal{FT} + \mathcal{FT}_{LID}$	82.4	73.2	77.8	**87.3**	62.9	**67.6**	**72.6**

Table 4: Performance comparison across various models. Our proposed methods are highlighted in gray. The best values in each column are in **bold**.