Avoiding Shortcut Learning in Deep Models for Enhanced Defense Against Adversarial Attacks

Ambroise LAROYE
ENSEA, ETIS
Cergy, FRANCE
ambroise.laroye–langouet@ensea.fr

Son VU
ENSEA, ETIS
Cergy, FRANCE
son.vu@ensea.fr

***Abstract*—Recently, Large Language Models (LLMs) have made significant strides in text classification. However, they remain susceptible to spurious correlations or shortcuts—unintended associations between training data and task labels—that can hinder generalization and adversarial robustness. Most existing approaches address this issue by identifying a limited set of task-specific shortcuts using human priors or data augmentation, both of which require substantial expertise and manual effort. In this project, we explore solutions to mitigate shortcut learning. Our approach revolves around a two-phase fine-tuning strategy: first, an initial fine-tuning on the full dataset, followed by a second fine-tuning phase on a subset identified as spurious using one of three extraction techniques—forgettable examples, attention scores, or LID scores. We evaluate this method on four classification datasets (MNLI, HANS, FEVER, and FEVER-Symmetric) and find that this informed fine-tuning strategy significantly improves robustness. Among the extraction methods, importance scores prove to be the most effective in identifying spurious correlations. Notably, using a BERT-base model, we achieve an accuracy of 73.6% on HANS, outperforming the 62.9% obtained with standard fine-tuning and surpassing all previous studies.**[1]

I. Introduction

This project report traces the chronological evolution of my work: the methodology I followed, the choices I had to make, the challenges I encountered, and how I addressed them. I will also present my results, detailing the experiments conducted and the methods used to obtain them. Finally, I will analyze the relevance and significance of the findings.
But first, let's return to the main topic, "spurious correlations". A thorough understanding of the subject, its challenges, and the current state of research is essential before delving deeper into the study.

A. Subject

First, let's break down the subject: *Avoiding Shortcut Learning in Deep Models for Enhanced Defense Against Adversarial Attacks*. Given its broad scope, I had to quickly choose a specific research direction. But what should I focus on—Computer Vision, NLP...?
I found that the topic is widely explored in both fields. Arbitrarily, I chose NLP, mainly because it presents numerous aspects to investigate. As I read first papers, I quickly realized that there is still much work to be done in this area.
My approach to this project began with gaining a deep understanding of the topic. One of my initial mistakes was focusing solely on keywords, losing sight of the overall context. For instance, I became too fixated on the term "attack", whereas it is actually a consequence of the problem rather than the core issue itself. The subject itself focuses on how can we avoid shortcut learning ? Is there a method that can increase models robustness to spurious correlations ?
Next, I conducted a state of the art to identify work that have already been treated, studies that work or not and potential research directions. This helped me refine my focus and formulate a specific research question that I will explore and develop further.

B. Context and Motivation

Nowadays, shortcut learning is a major challenge in deep learning models. **Shortcut learning** refers to models relying on superficial patterns in data, such as background features or syntactic heuristics, which can lead to significant errors when tested on diverse or unexpected data. This not only limits model performance but also creates vulnerabilities that adversaries can exploit through adversarial attacks. Adversarial attacks, exploit these vulnerabilities to manipulate model predictions or create deceptive content. The project aims to develop optimization-based learning techniques to mitigate these issues, improving the robustness and security of AI systems.
I propose the following definition : 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. 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 [1, 2]. 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 [3].
LLMs, such as BERT [4] and RoBERTa [5], have achieved strong performance in natural language inference (NLI) tasks but remain vulnerable to spurious correlations [6, 7].

[1]The code is available at https://github.com/Ambroise012/robustness_to_spurious_correlation.

II. State of the Art

To enhance robustness against spurious correlations, several methods have been explored. These methods can be categorized into two main approaches: (1) those that focus on identifying and removing spurious correlations to improve robustness and (2) those that address these correlations by mitigating their impact. A crucial first step is defining a method to identify spurious correlations before applying mitigation strategies.

A. Identification of Spurious Correlations

Identifying spurious correlations in data is a key challenge. Several approaches have been proposed:

- **Supervised Identification**:
  - [3] introduce a method based on treatment effect estimators to detect spurious correlations at the word level.
  - However, this approach relies on human annotation, limiting its scalability.
- **Automated Identification**:
  - [8] propose an automated framework that leverages attention scores and integrated gradients.
  - This method differentiates genuine correlations from spurious ones through cross-dataset analysis, eliminating the need for manual annotation.
  - Another approach is **example forgetting** [9], where instances frequently misclassified or later forgotten during training highlight hard-to-learn patterns and minority cases.
  - [10] propose a **two-stage fine-tuning** method:
    1) Training the model on the full dataset.
    2) Fine-tuning it exclusively on minority examples.

B. Mitigation Strategies

Once identified, spurious correlations can be mitigated through various learning strategies:

- **Pre-trained Language Models (PLMs)**:
  - PLMs improve robustness by generalizing from minority examples [11].
  - Multi-task learning further enhances performance by incorporating auxiliary tasks [12].
  - Other works suggest adding richer contextual information or revising fine-tuning strategies [13].
- **Transfer Learning and Adversarial Learning**:
  - Transfer learning enables models to generalize better to diverse inputs.
  - Adversarial learning exposes models to challenging examples, improving their robustness [14].
  - Benchmark datasets like HANS assess a model's reliance on heuristics [15].

Our intuition is that combining the approach proposed by [10], which involves PLMs and fine-tuning, with a more effective identification method for spurious correlations could further improve results. In particular, the example-forgetting approach does not account for all types of spurious correlations. Specifically, some spurious correlations may persist as consistent misclassifications without the model ever revising its predictions. This limitation underscores the need for a more comprehensive mitigation strategy.

III. Methodology

Our approach is a two-stage process outlined in Figure 2.

- 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 disproportionality or overfitting of any single class.
To identify spurious correlation features, our study consider two methods :
We propose to identify spurious correlation through **important score**. Defining importance score for an example as:

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

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 1 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** [16] : 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.

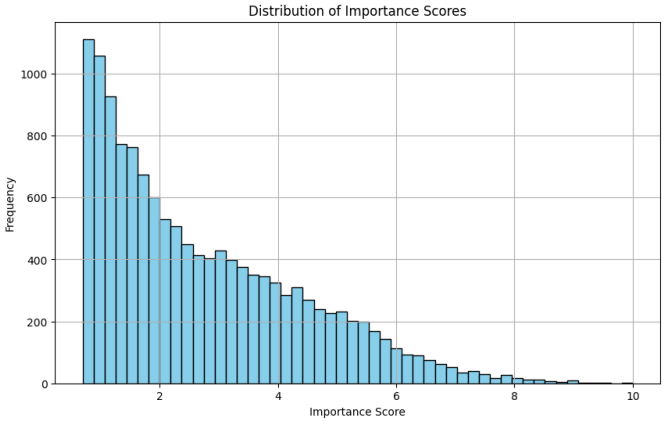

Fig. 1: Distribution of important scores for token selection.

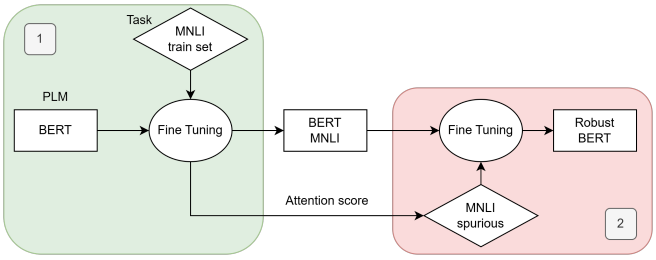

Fig. 2: Two step fine tuning method on MNLI task with identification of spurious correlation through important score

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.

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.

Model This study focuses on BERT model [4]. I choose BERT for its major use in previous work [10], [17] [18] [19]. We used the pre-trained *googlebert/bert-base-cased*[2] model with 12 layers and 12 attention heads.

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 [20], 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) [15], 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 [21], 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 [22], 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.

[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 I: Models accuracy on MNLI and HANS datasets

IV. Experiment 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 I 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 I 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.

In Table II, 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.

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

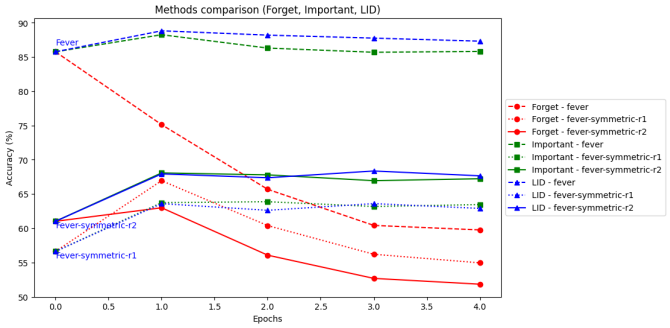

Fig. 3: Accuracy comparison of models fine-tuned with different extraction methods across FEVER datasets.

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 VII). Moreover, the proposed method is both easy to implement and computationally efficient.

A. Improvements

The results obtained are very promising. However, when considering writing a research paper, we felt it was important to add another contribution, particularly in the fine-tuning process.

1) Fine Tuning: One initial approach was to incorporate Explanation-based Finetuning as proposed by [23].
I first implemented the Explanation-based Finetuning method. However, their study is based on tasks such as ComVE (+1.2), CREAK (+9.1), e-SNLI (+15.4), and SBIC (+6.5), which focus more on reasoning and plausibility, making them particularly well-suited for this approach. In contrast, our work focuses on MNLI and HANS—classification tasks involving entailment, contradiction, or neutrality—where generating meaningful ”explanations” proved to be quite challenging. Unfortunately, I was unable to obtain usable results.
I also explored other approaches to improve fine-tuning, including adversarial training—specifically, FreeLB [24]. However, this method also yielded inconclusive results, with significantly lower accuracy: 35% on MNLI and 50% on HANS.
Although these additional methods did not improve our study, it may be more productive to focus on analyzing our current results more deeply. Can we push the analysis further?

2) TextAttack: We employ *TextAttack* [25], a framework designed for generating adversarial examples on natural language models. The attack method modifies input text until a misclassification occurs, and we measure the number of queries, as well as the percentage of perturbed words required for a successful attack. The results of the adversarial attacks on both models are summarized in Table III.

From the results, we observe that our model exhibits a significantly lower perturbation rate (4.18%) compared to the original BERT (11.48%). This indicates that fewer word modifications are required to maintain correct predictions, suggesting greater resilience to small input changes.
However, the number of queries required to generate an adversarial example is slightly lower for the $\mathcal{FT} + \mathcal{FT}_{important}$ (52.5) than for the original BERT (63.0). This suggests that adversarial examples may be found more efficiently for our model.
The lower perturbation rate of the modified BERT suggests that it is more robust to adversarial attacks, as adversaries need to change fewer words to disrupt predictions.

Figure 3 presents a comparison of three extraction methods: forget events [9], 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.

3) Which label is the most spurious: 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.

4) Threshold: We can draw a parallel with 1. In fact, when varying the threshold for selected importance scores, accuracy does not change significantly. This is because the number of spurious examples retrained remains constant. Moreover, discussing a threshold in our case is not particularly mean-

Model	Perturbed Words (%)	Words/Input	Queries
$\mathcal{FT}$	11.48	30.8	63.0
$\mathcal{FT} + \mathcal{FT}_{important}$	4.18	30.8	52.5

TABLE III: Adversarial attack results comparing BERT (original) and BERT (modified).

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

TABLE IV: Number of spurious examples extracted per method and their label distribution for MNLI

Model	MNLI	HANS	Threshold
BERT	**84.2**	62.9	/
$BERT + Forget_{BERT}$	82.9	69.0	/
$BERT + Important_{BERT}$	82.4	73.3	0.08
$BERT + Important_{BERT}$	82.7	**73.6**	0.1
$BERT + Important_{BERT}$	82.5	73.4	0.3

TABLE V: Accuracy after 4 epochs on BERT ; BERT + Forgettable examples ;
BERT + Important samples

ingful, as most of the essential information is concentrated near or equal to zero (over 80%). Indeed, a large number of words, such as "the", "a", or "I", do not need to be retrained. Including them would introduce noise into the dataset rather than improving the model's robustness.

5) Who is spurious ?: In this part we In the following table, we give few examples of sentences that have been classified into spurious correlation with our extracting methods: These examples may reflect spurious correlations for several reasons:

Lexical Overlap

E.g.: "All this is their information again." vs. "This information belongs to them."

The high lexical overlap—"their," "them"—may lead a model to predict entailment even if the meanings are not strictly equivalent.

E.g.: "What about me?" vs. "Me too?"

Despite minimal content, the model might infer entailment due to pronoun similarity, though the intent subtly differs.

Negation Cues Leading to Contradiction Bias

E.g.: "Turned out, I wasn't completely wrong." vs. "I was 100 percent wrong."

The presence of the word "wrong" alongside negation could trigger contradiction predictions.

Paraphrasing

E.g.: "Get individuals to invest their time and the funding will follow." vs. "If individuals invest their time, funding will come along, too."

The syntactic rephrasing might trigger entailment based on surface-level similarity, even though the causal nuance may differ subtly.

Spurious Associations

E.g.: "Joe Montana and John Elway..." vs. "John Elway and Joe Montana played football together.";

The model may rely on the co-occurrence of names rather than semantic content, potentially leading to a neutral label despite conflicting timelines or contexts.

Ambiguous Phrasing

E.g.: "That could be it." vs. "No, that couldn't be."

Spurious words

This time, we focus on certain words that might frequently appear in sentences classified as spurious examples. We are wondering if there are any 'spurious words.'

Therefore, we will examine the frequency of the most common words in the spurious examples and compare this frequency to the occurrence of the same words in the original MNLI dataset.

For example, if the word 'the' is the most frequent among the spurious examples, this does not make it a 'spurious word' since it is also very common in the MNLI dataset.

We are thus looking for words that are at least twice as frequent in the spurious examples compared to the original dataset (with frequency normalized by the total number of examples in the spurious set and the base dataset). For example, in the list of words above, we can interpret the potential reasons for their misclassification: 'Densities', 'graph', and 'liters' may be frequent in scientific contexts, often linked to neutral hypotheses due to the nature of the selected examples. 'Fetus', 'compost', and 'bullfights' are associated with sensitive or controversial topics, which could relate them to contradiction labels if the dataset includes debates on these subjects.

Names like 'Yellowstone', 'Hideyoshi', 'Amenophis', 'Winnie', 'Dowd' or historical figures like 'Hideyoshi' may frequently appear with specific labels due to their context in the corpus.

Words such as 'skimmers', 'invalides', 'townhouse', and 'beneficiary' are rare. Models tend to overfit these words because their low frequency is strongly associated with specific labels in the available examples.

V. Discussion

In our study, we implemented an experiment using a two-step fine-tuning approach. By specifically addressing the identification of spurious correlations, we significantly improved the model's robustness. Indeed, our new results largely surpass those obtained in previous experiments, as demonstrated in Table VII. However, despite this improvement, we were unable to push the performance beyond a certain threshold.

To gain deeper insights into the behavior of our model, we conducted a thorough analysis of the results.

In a nutshell, while our fine-tuning and correction strategies have led to a significant improvement in robustness, alternative methods—such as attention-based analysis or feature importance visualization—may be more suitable for verifying the effectiveness of spurious correlation mitigation.

Sentence 1	Sentence 2	Label
How do you know? All this is their information again.	This information belongs to them.	entailment
Take a remarkable statistic that Shesol cites but lets pass relatively unexamined.	They had data that was very relevant but under used.	entailment
Get individuals to invest their time and the funding will follow.	If individuals will invest their time, funding will come along, too.	entailment
well because how hot i mean like like in the coldest that it gets in winter down there how much is it	It's hot all the time where I live, including winter.	neutral
The man should have died instantly.	The man was perfectly fine.	contradiction
In summer the rice forms a green velvety blanket, then turns golden in autumn when it ripens and is harvested.	The rice is golden and harvestable in the summer, but turns green in autumn.	contradiction
well see that isn't too bad a couple hours	I don't like waiting, not even for 10 minutes.	contradiction
Less loud.	Please be quiet.	entailment
Turned out, I wasn't completely wrong.	I was 100 percent wrong.	contradiction
He'd stopped wondering and now accepted; he meant to get away from here at the first chance and he was somehow sure he could.	The doctors office was a terrible place and he wanted out.	neutral
The Inglethorps did not appear.	The Inglethorps were the first ones to turn up.	contradiction
equivalent to increasing national saving to 19.	National savings are 18 now.	neutral
What about me?	Me too?	entailment
they use the the injection thing or whatever it is	They use lethal injection.	entailment
Stop blaming John.	Quit holding John responsible.	entailment

TABLE VI: Sentence Pairs with Corresponding Labels that may lead to hallucinations

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 [2]	83.0	63.6	73.3	87.6	62.2	68.2	72.7
Post-HOC, argmin [2]	81.0	67.2	74.1	85.3	61.8	67.4	71.5
Reweighting [22]	81.4	68.6	75.0	84.6	61.7	64.9	70.4
PoE [17, 26]	84.2	64.6	74.4	82.3	62.0	64.3	69.5
Reg-conf [18]	**84.3**	69.1	76.7	86.4	60.5	66.2	69.2
SoftLE [19]	81.2	68.1	74.6	87.5	60.3	66.9	71.6
$\mathcal{FT} + \mathcal{FT}_{forget}$ [10]	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 VII: Performance comparison across various models. Our proposed methods are highlighted in gray. The best values in each column are in **bold**.

VI. Conclusion

This project has been a valuable experience in the world of research, providing me with insights into the methodologies and challenges involved. I have come to realize that conducting research is far from easy, requiring rigorous experimentation and critical analysis.

Current challenges in LLMs remain vast, and the issue of spurious correlations appears to be relatively underexplored, despite its significant impact on security and model robustness. Addressing these correlations can lead to substantial improvements in performance and reliability, making it a crucial area of study.

Throughout this project, I have gained a deep understanding of spurious correlations in NLP, explored existing research, and learned about various mitigation techniques. The results obtained are quite satisfying, and this experience has motivated me to write my first research paper. This endeavor has given me a deeper appreciation for the research process and the effort required to contribute to scientific advancements.

Overall, this project has been both challenging and rewarding, reinforcing my interest in research and pushing me to further explore this field.

VII. Acknowledgment

I would like to express my gratitude to Son VU for proposing this research topic and for guiding me throughout the project. I also extend my thanks to my school ENSEA and the ETIS laboratory for providing the necessary resources and a supportive research environment.

References

[1] M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour, "Causally motivated

Forget

Word	Spurious_Frequency	Normalized_Frequency
dutchman	11	5.597563165709285
workfare	11	4.867446231051552
shepard	12	4.361737531721521
extremists	11	3.99825940407806
pataki	12	3.9396338996194378
shoemaker	13	3.891354072418219
liters	12	3.4893900253772165
yellowstone	11	3.392462524672294
modesty	17	3.327222860736288
suckers	12	3.3007743483297993
jahangir	12	3.3007743483297993
dishwasher	11	3.2926842151231086
unlv	19	3.2774637950223853
stepson	12	3.131503868928271
hose	12	3.131503868928271
eastwood	12	3.131503868928271
mcgrath	11	3.109757314282936
stormed	12	3.0532162722050646
postwar	17	3.0353612062857365

Attention

Word	Spurious_Frequency	Normalized_Frequency
liters	11	4.580334699022354
townhouse	15	3.8352084800426414
lifers	12	3.801858841085749
chan	14	3.517811916177043
bullfights	12	3.497710133798889
dowd	21	3.188017049035446
invalides	11	3.143366950309459
warming	19	3.0766894695453186
gifted	19	3.042879695154711
beneficiary	11	3.0247493295430643
winnie	12	3.01526735672318
stanford	12	2.9147584448324073
compost	12	2.9147584448324073
fetus	14	2.8737055089896972
hideyoshi	14	2.833792932475951
densities	12	2.8207339788700714
graph	11	2.812486218697937
amenophis	14	2.7949738512091575
skimmers	14	2.7949738512091575

Fig. 4: Examples of words classified as spurious are those that appear at least 10 times in spurious examples and have a frequency at least twice as high in spurious examples as in the MNLI dataset

shortcut removal using auxiliary labels," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 739–766, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v151/makar22a.html

[2] U. Honda, T. Oka, P. Zhang, and M. Mita, "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*, vol. 12, pp. 1268–1289, Oct. 2024, arXiv:2406.12060 [cs]. [Online]. Available: http://arxiv.org/abs/2406.12060

[3] Z. Wang and A. Culotta, "Identifying Spurious Correlations for Robust Text Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3431–3440. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.308/

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019, arXiv:1810.04805 [cs]. [Online]. Available: http://arxiv.org/abs/1810.04805

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, arXiv:1907.11692 [cs]. [Online]. Available: http://arxiv.org/abs/1907.11692

[6] W. Ye, G. Zheng, X. Cao, Y. Ma, and A. Zhang, "Spurious Correlations in Machine Learning: A Survey," May 2024, arXiv:2402.12715 [cs]. [Online]. Available: http://arxiv.org/abs/2402.12715

[7] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut Learning in Deep Neural Networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, Nov. 2020, arXiv:2004.07780 [cs]. [Online]. Available: http://arxiv.org/abs/2004.07780

[8] T. Wang, R. Sridhar, D. Yang, and X. Wang, "Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models," May 2022, arXiv:2110.07736 [cs]. [Online]. Available: http://arxiv.org/abs/2110.07736

[9] M. Toneva, A. Sordoni, R. T. d. Combes, A. Trischler, Y. Bengio, and G. J. Gordon, "An Empirical Study of Example Forgetting during Deep Neural Network Learning," Nov. 2019, arXiv:1812.05159 [cs]. [Online]. Available: http://arxiv.org/abs/1812.05159

[10] Y. Yaghoobzadeh, S. Mehri, R. Tachet des Combes, T. J. Hazen, and A. Sordoni, "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, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 3319–3332. [Online]. Available: https://aclanthology.org/2021.eacl-main.291/

[11] L. Tu, G. Lalwani, S. Gella, and H. He, "An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 621–633, 2020, place: Cambridge, MA Publisher: MIT Press. [Online]. Available: https://aclanthology.org/2020.tacl-1.40/

[12] J. Yu, Y. Zhou, Y. He, N. L. Zhang, and R. Silva, "Fine-Tuning Pre-trained Language Models for Robust Causal Representation Learning," Oct. 2024, arXiv:2410.14375 [cs]. [Online]. Available: http://arxiv.org/abs/2410.14375

[13] R. Schwartz and G. Stanovsky, "On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations," Apr. 2022, arXiv:2204.12708 [cs]. [Online]. Available: http://arxiv.org/abs/2204.12708

[14] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A New Benchmark for Natural Language Understanding," May 2020, arXiv:1910.14599 [cs]. [Online]. Available: http://arxiv.org/abs/1910.14599

[15] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference," Jun. 2019, arXiv:1902.01007 [cs]. [Online]. Available: http://arxiv.org/abs/1902.01007

[16] M. Savić, V. Kurbalija, and M. Radovanović, "Local Intrinsic Dimensionality Measures for Graphs, with Applications to Graph Embeddings," Jul. 2023, arXiv:2208.11986 [cs]. [Online]. Available: http://arxiv.org/abs/2208.11986

[17] C. Clark, M. Yatskar, and L. Zettlemoyer, "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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4069–4082. [Online]. Available: https://aclanthology.org/D19-1418/

[18] P. A. Utama, N. S. Moosavi, and I. Gurevych, "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*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8717–8729. [Online]. Available: https://aclanthology.org/2020.acl-main.770/

[19] Z. He, H. Deng, H. Zhao, N. Liu, and M. Du, "Mitigating Shortcuts in Language Models with Soft Label Encoding," Sep. 2023, arXiv:2309.09380 [cs]. [Online]. Available: http://arxiv.org/abs/2309.09380

[20] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," Feb. 2018, arXiv:1704.05426 [cs]. [Online]. Available: http://arxiv.org/abs/1704.05426

[21] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "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)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 809–819. [Online]. Available: https://aclanthology.org/N18-1074/

[22] T. Schuster, D. Shah, Y. J. S. Yeo, D. Roberto Filizzola Ortiz, E. Santus, and R. Barzilay, "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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3419–3425. [Online]. Available: https://aclanthology.org/D19-1341/

[23] J. M. Ludan, Y. Meng, T. Nguyen, S. Shah, Q. Lyu, M. Apidianaki, and C. Callison-Burch, "Explanation-based Finetuning Makes Models More Robust to Spurious Cues," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4420–4441. [Online]. Available: https://aclanthology.org/2023.acl-long.242/

[24] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, "FreeLB: Enhanced Adversarial Training for Natural Language Understanding," Apr. 2020, arXiv:1909.11764 [cs]. [Online]. Available: http://arxiv.org/abs/1909.11764

[25] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 119–126. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.16/

[26] R. Karimi Mahabadi, Y. Belinkov, and J. Henderson, "End-to-End Bias Mitigation by Modelling Biases in Corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 8706–8716. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.769