Robustness Challenges of Large Language Models in Natural Language Understanding: A Survey

Mengnan Du[1], Fengxiang He[2], Na Zou[1], Dacheng Tao[2], and Xia Hu[3]
[1]Texas A&M University
[2]JD Explore Academy
[3]Rice University
dumengnan@tamu.edu, hefengxiang@jd.com, nzou1@tamu.edu
dacheng.tao@gmail.com, xia.hu@rice.edu

Abstract

Large language models (LLMs) have achieved state-of-the-art performance on a series of natural language understanding tasks. However, these LLMs might rely on dataset bias and artifacts as shortcuts for prediction. This has significantly hurt their Out-of-Distribution (OOD) generalization and adversarial robustness. In this paper, we provide a review of recent developments that address the robustness challenge of LLMs. We first introduce the concepts and robustness challenge of LLMs. We then introduce methods to identify shortcut learning behavior in LLMs, characterize the reasons for shortcut learning, as well as introduce mitigation solutions. Finally, we identify key challenges and introduce the connections of this line of research to other directions.

1 Introduction

Large language models (LLMs), such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), have achieved state-of-the-art performance in a series of high-level natural language understanding (NLU) tasks, such as natural language inference (NLI), question answering (QA), etc. However, the superior performance has only been observed in the benchmark test data that have the same distribution as the training set. Recent studies indicate that these LLMs are not robust and that the models do not remain predictive when the distribution of inputs changes (Niven and Kao, 2019; Utama et al., 2020b; Du et al., 2021a). Specifically, these LLMs have low generalization performance when applied to out-of-distribution (OOD) test data and are also vulnerable to various types of adversarial attack.

A major reason for the low robustness of LLMs is **shortcut learning**. The shortcut learning behavior has also been called other names in the literature, such as *learning bias, superficial correlations, Clever Hans effect*, etc. (Heinzerling, 2019; Lapuschkin et al., 2019). The shortcut learning behavior has been observed for a series of NLU tasks. For example, recent empirical analysis indicates that the performance of BERT-like models for the NLI task could be mainly explained by relying on spurious statistical cues such as unigrams ‘not’, ‘do’, ‘is’ and bigrams ‘will not’ (Niven and Kao, 2019; Gururangan et al., 2018). Similarly, for the reading comprehension task, the models rely on the lexical matching of words between the question and the original passage, while ignoring the designed reading comprehension task (Lai et al., 2021). The current standard approach to training LLM is using empirical risk minimization (ERM) on NLU datasets that typically contain various types of artifacts and biases. As such, LLMs have learned to rely on dataset artifacts and biases and capture their spurious correlations with certain class labels as shortcuts for prediction. Shortcut learning has significantly hurt the models’ robustness, thus attracting increasing attention from the NLP community to address this issue.

In this work, we provide a review of the shortcut learning problem in LLMs, including its concept and robustness challenges in Section 2, detection approaches in Section 3, characterization of the corresponding reasons in Section 4, and mitigation approaches in Section 5. We also provide a further discussion of future research directions and connection with other directions in Sections 6 and 7. Note that in this work we mainly focus on the widely used pre-training and fine-tuning paradigm of LLMs in NLU tasks (see Figure 2).

2 Shortcut Learning Phenomena

2.1 What is Shortcut Learning?

Features captured by the model can be broadly categorized as useless features, robust features, and non-robust features (Ilyas et al., 2019) (see Figure 1). Shortcut learning refers to the phenomenon

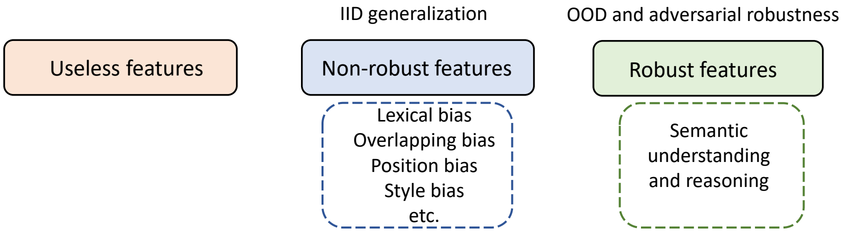

Figure 1: Features can be generally grouped into useless features, robust features, and non-robust features. Non-robust features indicate various kinds of biases captured by the model, which are not robust in the OOD setting. In contrast, robust features denote features of high-level semantic understanding that are robust to changes in the input.

that LLMs (especially those trained with standard ERM-based method) highly rely on non-robust features as shortcuts, failing to learn robust features and capture high-level semantic understanding and reasoning. Non-robust features do help generalization for development and test sets that share the same distribution with training data. However, they cannot generalize to OOD test sets and are vulnerable to adversarial attacks. Non-robust features are oriented from biases in the training data and come in different formats. In the following, we introduce several representative ones.

- *Lexical Bias*: Some lexical features have a high correlation of co-occurrence with certain class labels. These lexical features mainly consist of low-level functional words such as stop words, numbers, negation words, etc. One typical example is the NLI task, where LLMs are highly dependent on unintended lexical features to make predictions (Niven and Kao, 2019; Du et al., 2021a). For example, these models tend to give contradiction predictions whenever there exist negation words, e.g., 'never', 'no', in the input samples.
- *Overlap Bias*: It occurs in NLU applications with two branches of text, e.g. NLI, QA, and reading comprehension (Zhou and Bansal, 2020; McCoy et al., 2019). LLMs use the overlap of features between the two branches of inputs as spurious correlations as shortcuts. For example, reading comprehension models use the overlap between the passage and question pair for prediction rather than solving the underlying task (Lai et al., 2021). Similarly, QA models excel at test sets by relying on heuristics of question and context overlap (Sen and Saffari, 2020).
- *Position Bias*: The distribution of the answer positions may be highly skewed in the training set for some applications. The LLMs would predict answers based on spurious positional cues. Take the QA task for example, the answers lie only in the k-th sentence of each passage (Ko et al., 2020). As a result, QA models rely on this spurious cue when predicting answers.
- *Style Bias*: Text style is a kind of pattern that is independent of semantics (DiMarco and Hirst, 1993). Models have learned to rely on the spurious text style as shortcuts, which can be further utilized for adversarial attacks (Qi et al., 2021).

2.2 Generalization and Robustness Challenge

The shortcut learning behavior could significantly hurt LLMs' **OOD generalization** as well as **adversarial robustness**. First, shortcut learning could lead to a high degradation of performance for OOD data. A common assumption is that training and test data are independently and identically distributed (IID). This IID assumption will not hold when LLMs are deployed in real-world applications that exist distribution shifts. These data typically do not contain the same types of bias and artifacts as the training data (Koh et al., 2021).

$$\begin{aligned}&\text{IID: } \boldsymbol{P}_{train}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{P}_{test}(\boldsymbol{X},\boldsymbol{Y})\\&\text{OOD: } \boldsymbol{P}_{train}(\boldsymbol{X},\boldsymbol{Y}) \neq \boldsymbol{P}_{test}(\boldsymbol{X},\boldsymbol{Y})\end{aligned} \tag{1}$$

Taking BERT-base for example, there is a reduction in accuracy of more than 20% on the OOD test set, compared to the accuracy on the in-distribution test sets for three NLU tasks (Du et al., 2021b). To some extent, these models have solved the dataset rather than the underlying task. Second, shortcut learning also results in models that could be easily fooled by adversarial samples, when small and often imperceptible human-crafted perturbations are added to the normal input (Wang et al., 2021a). One typical example is for the multiple choice reading comprehension task (Si et al., 2019). BERT models are attacked by adding distracting information, resulting in a significant performance drop. Further analysis indicates that these models are highly driven by superficial patterns, which inevitably leads to their adversarial vulnerability.

3 Identification of Shortcut Learning

In this section, we discuss methods to identify shortcut learning problems in NLU models.

3.1 Comprehensive Performance Testing

Traditional evaluations employ IID training-test split of data (Agrawal et al., 2018). The test sets are drawn from the same distribution as the training sets and thus share the same kind of biases as the training data. Models that simply rely on memorizing superficial patterns could perform acceptably on the IID test set. This type of evaluation has failed to identify the shortcut learning problem. Therefore, it is desirable to perform more comprehensive tests beyond the traditional IID testing.

First, the OOD generalization test has been proposed as an alternative to the IID test. Take MNLI (Williams et al., 2018) for example, the HANS evaluation set is proposed to evaluate whether NLI models have syntactic heuristics: the lexical overlap heuristic, the subsequence heuristic, and the constituent heuristic (McCoy et al., 2019). Similarly, for the FEVER fact verification task (Thorne et al., 2018), a symmetric test set is constructed that shares a philosophy similar to HANS (Schuster et al., 2019). These OOD tests have revealed dramatic performance degradation and exposed the shortcut learning problem of state-of-the-art LLMs.

Second, adversarial attacks could also be implemented to test the robustness of LLMs. For example, adversarial attacks have been used to reveal statistical bias in machine reading comprehension models (Lai et al., 2021). The adversarial examples created through TextFooler (Jin et al., 2020) are used to test the generalization of common sense reasoning models (Branco et al., 2021). The results indicate that the models have learned non-robust features and fail to generalize towards the main tasks associated with the datasets.

Third, randomization ablation methods are proposed to analyze whether LLMs have used these essential factors to achieve effective language understanding. For example, word order is a representative one among these significant factors. Recent ablation results indicate that word order does not matter for pre-trained language models (Sinha et al., 2021). In particular, LLMs are pre-trained first on sentences with randomly shuffled word order and then fine-tuned on various downstream tasks. The results show that these models still achieve high accuracy. Similarly, another study (Pham et al., 2020) has observed that LLMs are insensitive to word order in a wide set of tasks, including the entire GLUE benchmark (Wang et al., 2019). These experiments indicate that LLMs have ignored the syntax when performing downstream tasks, and their success can almost be explained by their ability to model higher-order word co-occurrence statistics.

3.2 Explainability Analysis

DNN explainability is another effective tool that the community has used to identify the shortcut learning problem. LLMs are usually considered black boxes, as their decision-making process is opaque and difficult for humans to understand. This presents challenges in identifying whether these models make decisions based on justified reasons or on superficial patterns. Explainability enables us to diagnose spurious patterns captured by LLMs.

The existing literature mainly employs the explanation in the format of feature attribution to analyze shortcut learning behavior in NLU models (Wang et al., 2021c; Du et al., 2021a). Feature attribution is the most representative paradigm among all explainability-based methods. In particular, for each token x_i within a specific input x, the feature attribution algorithm ψ will calculate the contribution score ψ_i, which denotes the contribution score of that token for model prediction. For example, the Integrated Gradient (Sundararajan et al., 2017) interpretation method is used to analyze the model behavior of BERT-based models (Du et al., 2021a). It is observed that LLMs rely on dataset artifacts and biases within the hypothesis sentence for prediction, including functional words, negation words, etc. (Du et al., 2021a). This shortcut learning behavior is summarized further using the long-tailed phenomenon. Specifically, the tokens in the training set could be modeled using a long-tailed distribution. The LLM models concentrate mainly on information on the head of the distribution, which typically corresponds to non-generalizable shortcut tokens. In contrast, the tail of the distribution is poorly learned, although it contains abundant information for an NLU task.

Beyond feature attribution, other types of explainability methods have also been used to analyze shortcut learning behaviors (Han and Tsvetkov, 2021). For example, instance attribution methods have been used to explain model prediction by identifying influential training data, which can be used

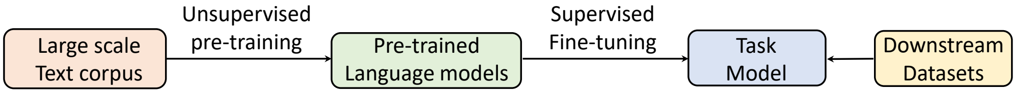

Figure 2: The pre-training then fine-tuning training paradigm. Shortcut learning can be attributed to different factors in this pipeline, including pre-trained language models, fine-tuning process, and downstream tasks.

to explain decision making logic for the current sample of interest (Han et al., 2020b). Empirical analysis indicates that the most influential training data share similar artifacts, e.g., high overlap between the premise and hypothesis for the NLI task. Furthermore, hybrid methods that combine feature attribution and instance attribution have also been used to identify artifacts in the data (Pezeshkpour et al., 2021). The resulting explanations have provided a more comprehensive perspective on the shortcut learning behavior of LLMs.

4 Origins of Shortcut Learning

The issue of learning shortcuts in NLU models originates from different factors in the training pipeline (see Figure 2). In this section, we analyze these reasons, particularly focusing on the following three factors: training datasets, LLM model, and the fine-tuning training process.

4.1 Skewed Training Dataset

From the data perspective, the shortcut learning of the NLU models can be traced to a large extent to the annotation artifacts and collection artifacts of the training data (here we mainly refer to the downstream datasets in Figure 2). Training sets are typically built through the crowd-sourcing process, which has the advantage of being low-cost and scalable. However, the crowd-sourcing process results in collection artifacts, where the training data are imbalanced with respect to features and class labels. Furthermore, when crowd workers author parts of the samples, they produce certain patterns of artifacts, i.e. annotation artifacts (Gururangan et al., 2018). Taking NLI as an example, the average sentence length of the hypothesis branch is shorter for the entailment category compared to the neutral category (Gururangan et al., 2018). This suggests that crowd workers tend to remove words from the premise to create a hypothesis for the entailment category, leading to the overlap bias in the training data. Models trained on the skewed datasets will capture these artifacts and even amplify them during inference time.

4.2 LLMs Models

The robustness of NLU models is highly relevant to the pre-finetuned LLMs models. In particular, there are two key factors: model sizes (measured by the number of parameters) and pre-training objectives.

First, models with the same kind of architectures and pre-training objective but with different sizes could have significantly different generalization ability. It is shown that increasing the size of the model could lead to an increase in the representation power and generalization ability. From the empirical perspective, comparisons have been made between LLMs of different sizes but with the same architecture, e.g., BERT-base with BERT-large, RoBERTa-base with RoBERTa-large (Tu et al., 2020; Bhargava et al., 2021). The results indicate that the larger versions generally generalize consistently better than the base versions, with a significantly smaller accuracy gap between the OOD and IID test data. The smaller models have fewer parameters than the larger model and their model capacity is smaller. Therefore, smaller models are more prone to capture spurious patterns and are more dependent on data artifacts for prediction (Sanh et al., 2021). Another work (Du et al., 2021b) studies the impact of model compression on the robustness and generalizability of LLMs and finds that compressed LLMs are significantly less robust compared to their uncompressed counterparts. Compressed models with knowledge distillation have also been shown to be more vulnerable to adversarial attacks (Li et al., 2021). From a theoretical perspective, a recent analysis supports that there is a trade-off between the size of a model and its robustness, where large models tend to be more robust (Bubeck and Sellke, 2021).

Second, LLMs with similar model sizes but with different pre-training objectives also differ in the generalization ability. Here, we consider three kinds of LLMs: BERT, ELECTRA, and RoBERTa. BERT is trained with masked language modeling and next-sentence prediction. RoBERTa removes the next-sentence prediction from BERT and uses dynamic masking. ELECTRA is trained to distinguish between real input tokens and fake input tokens generated by another network. Empirical analysis shows that these three models have significantly different levels of robustness (Prasad et al., 2021). For the Adversarial NLI (ANLI) dataset (Nie et al., 2020), it is shown that ELECTRA and RoBERTa have significantly better performance than BERT, for both the base and the large versions. Similarly, another study has shown that RoBERTa-base outperforms BERT-base around 20% in terms of accuracy on the HANS test set (Bhargava et al., 2021). A possible reason is that different inductive biases are encoded by the models, since different architectures have distinct object functions during the pre-training stage.

4.3 Model Fine-tuning Process

The learning dynamics could reveal what knowledge has been learned during the course of model training. There are some observations. First, standard training procedures generally have a bias towards learning simple features (Shah et al., 2020). The models are based mainly on the simplest features and remain invariant to complex predictive features. Moreover, it has been observed that the models give overconfident predictions for easy samples and low-confidence predictions for hard samples. Second, models tend to learn non-robust and easy-to-learn features at the early stage of training (Hermann and Lampinen, 2020). For example, reading comprehension models have learned the shortcut in the first few training iterations, which has influenced further exploration of the models for more robust features (Lai et al., 2021). Third, longer fine-tuning could lead to better generalization. Specifically, a larger number of training epochs dramatically improves the generalizability of LLMs in NLU tasks (Tu et al., 2020).

The preference for non-robust features can be explained from two perspectives. First, current LLM training paradigms can be regarded as data-driven, corpus-based, statistical, and machine learning methods (Saba, 2021). It is argued that this data-driven paradigm might be useful in some NLP tasks, which however are not even relevant to NLU tasks that require high-level natural language understanding (Saba, 2021). Second, from the model optimization perspective, current practice uses gradient descent optimization. For instance, it has been theoretically proven that gradient descent methods tend to learn non-robust networks, by using a depth-2 ReLU network as an example (Vardi et al., 2022).

5 Mitigation of Shortcut Learning

In this section, we introduce approaches that alleviate the problem of shortcut learning. These methods are motivated mainly by the insights obtained in the last section. In particular, Section 5.1 introduces methods based on dataset refinement. The rest of the sections focus on model-centric mitigation approaches, typically augmenting the traditional ERM-based training paradigm with different degrees of prior knowledge, explicitly or implicitly suppressing the model from capturing non-robust features. Some mitigation methods require that the shortcuts be known a priori, while others assume that the shortcuts are unknown. These can be named **robust learning** methods, where the ultimate goal is to improve OOD generalization and adversarial robustness, while still exhibiting good predictive performance in IID datasets.

5.1 Dataset Refinement

Dataset refinement falls into the pre-processing mitigation family, with the aim of alleviating biases in the training datasets (Wu et al., 2022). First, when constructing new datasets, crowd workers will receive additional instructions to discourage the use of words that are highly indicative of annotation artifacts (Han et al., 2020a). Second, debiased datasets can also be developed by filtering out bias in existing data. For example, adversarial filtering is used to build a large-scale data set for the NLI task to reduce annotation artifacts that can be easily detected by a committee of strong baseline methods (Zellers et al., 2018). As a result, models trained on this dataset have to learn more generalizable features and rely on common sense reasoning to succeed. Third, we can also reorganize the train and test split, so that the bias distribution in the test set is different from that in the training set (Agrawal et al., 2018). Lastly, various kinds of data augmentation methods have been proposed. Representative examples include counterfactual data augmentation (Kaushik et al., 2020), mixup data augmentation (Si et al., 2021), syntactically informative example augmentation by applying syntactic transformations to sentences (Min et al., 2020), etc.

5.2 Adversarial Training

Adversarial training aims to learn better representations that do not contain information about artifacts or bias in the data. It is typically implemented in two ways in the NLP domain (Stacey et al., 2020; Rashid et al., 2021). First, the task classifier and adversarial classifier jointly share the same encoder (Stacey et al., 2020). The goal of the adversarial classifier is to provide the correct predictions for the artifacts in the training data. Then the encoder and task classifier can be trained to optimize the task objective while reducing the performance of the adversarial classifier in predicting artifacts. Second, adversarial examples are generated to maximize a loss function, and the model is trained to minimize the loss function. For example, the generator based on the masked language model is used to perturb the text to generate adversarial samples (Rashid et al., 2021). Despite the difference, both approaches leverage the MinMax formulation during the debiasing process.

5.3 Explanation Regularization

This category aims to regularize model training using prior knowledge established by humans (Liu and Avci, 2019; Han and Tsvetkov, 2021). Specifically, it is achieved by regularizing the feature attribution explanations with rationale annotations created by domain experts, to enforce the model to make the right predictions for the right reasons (Liu and Avci, 2019). These systems are trained to explicitly encourage the network to focus on features in the input that humans have annotated as important and suppress the models' attention to superficial patterns. For the NLI task, natural language explanations have been used to supervise the models, to encourage the model to pay more attention to the words present in the explanations (Stacey et al., 2022). It has significantly improved the models' OOD generalization performance. Note that this type of method can only be used when prior knowledge is known in advance about shortcuts.

5.4 Product-of-Expert (PoE)

The goal is to train a debiased model by training it as an ensemble with a bias-only model (Clark et al., 2019; He et al., 2019; Sanh et al., 2021). This paradigm usually contains two stages. In the first stage, a bias-only model is explicitly trained to capture the bias of the data set, e.g. the hypothesis-only bias for the NLI task. During the second stage, the debiased model will be trained using cross-entropy loss, by combining its output with the output of the bias-only model: $\hat{p}_i = \operatorname{softmax}\left(\log\left(p_i\right) + \log\left(b_i\right)\right)$. The parameters of the bias-only model are fixed during this stage, and only the debiased model parameters are updated by backpropagation. The goal is to encourage the debiased model to utilize orthogonal information with information from the bias-only model to make predictions.

5.5 Training Samples Reweighting

The main idea of reweighting is to place higher training weights on hard training samples, and vice versa (Schuster et al., 2019; Yaghoobzadeh et al., 2019; Utama et al., 2020b). It is also called *worst-group loss minimization* in some literature (Nam et al., 2020; Liu et al., 2021a). The underlying assumption is that improving the performance of the worst group (hard samples) is beneficial for model robustness. It is typically achieved through two-stage training. In the first stage, the weight indexing model is trained; and in the second stage, the predictions of the indexing model are used as weights to adjust the importance of a training instance. Both soft weights (Utama et al., 2020b) and hard weights (Liu et al., 2021a) could be used in the second stage. Another representative example is focal loss (Lin et al., 2017), which is based on a regularizer to assign higher weights to hard samples bearing less confident predictions.

5.6 Confidence Regularization

This mitigation scheme regularizes confidence in the model output, with the aim of encouraging the debiased model to give a higher uncertainty (lower confidence) for these biased samples. It is based on the observation that models tend to make overconfident predictions on biased examples (Utama et al., 2020a). This relies on the training of a bias-only model to quantify the degree of bias of each training sample. The debiasing process is typically achieved through the knowledge distillation framework. In the first stage, the biased teacher model is trained using standard ERM loss, and the bias degree obtained from the bias-only model will be used to rescale the output distribution of the teacher model. In the second stage, the smoothed confidence values of the teacher model can be used to guide the training of the debiased model.

5.7 Partitioning Data into Environments

This line of methods follows the principle of invariant risk minimization (Arjovsky et al., 2019), which encourages models to learn invariants in multiple environments. For example, training data has been partitioned into several non-IID subsets (i.e., training environments), where spurious correlations vary across environments and reliable ones remain stable across environments (Teney et al., 2020). The training scheme is designed to encourage the model to rely on stable correlations and suppress spurious correlations. Another work proposes an inter-environment matching objective by maximizing the inner product between gradients from different environments, with the goal of increasing model generalization (Shi et al., 2022).

5.8 Contrastive Learning

Contrastive learning can be used to guide the training of representations. The goal is to construct the instance discrimination task to guide the model to capture the robust and predictive features, while suppressing the undesirable non-robust features. The instance discrimination task should be carefully designed; otherwise, it is possible to suppress robust predictive features (Robinson et al., 2021).

6 Future Research Directions

Despite the progress introduced in the previous three sections, there are still many research challenges. In this section, we discuss the challenges that deserve further research from the community.

6.1 More Inductive Bias

It is suggested to introduce more inductive bias into the model architecture to improve robustness and generalization beyond IID benchmark datasets (Marasović, 2018). Recently, some work has begun to induce certain kinds of linguistic structure in neural architectures. For example, TableFormer is proposed for robust table understanding (Yang et al., 2022). It proposes a robust and structurally aware table-text encoding architecture, where tabular structural biases are incorporated through learnable attention biases. Note that inductive biases are highly task-dependent and should be carefully designed for each specific task to accommodate its unique characteristic.

6.2 Better Pre-training Objectives

It is desirable to invest more effort in designing better pre-training objectives to improve model robustness. Recent studies indicate that choosing a better pre-trained model could bring much better generalization performance than robust learning methods as introduced in Section 5. For example, RoBERTa-base with a standard fine-tuning loss could even outperform the BERT-base with robust learning objectives in terms of generalization performance on the HANS test set (Bhargava et al., 2021). This indicates the essential role of pre-training in NLU models' generalization performance and calls for more efforts from the community to improve the pre-trained language models.

6.3 Introducing More Domain Knowledge

NLU tasks might contain various types of bias, which are not fully known even by domain experts. This is distinct from the literature that works with the toy task (e.g., Colored MNIST (Arjovsky et al., 2019)), which typically contains a single type of bias and the bias is fully known. As such, most existing mitigation methods for NLU tasks rely on heuristics of human prior knowledge. Some examples include: i) weak models are more prone to capture biases, ii) non-robust models tend to give overconfident predictions for easy samples, etc. Unfortunately, this prior knowledge can only identify a limited number of biases existing in the data. Although it is possible to alleviate the usage of some identified shortcuts, models could use other shortcuts for prediction. This could explain why existing mitigation methods have only a limited improvement in generalization. Therefore, it is recommended to incorporate more human-like common sense knowledge in model training.

6.4 Analyzing Debiased Models

It is commonly believed that debiased algorithms achieve better generalization since they can learn more robust features, compared to biased models that rely mainly on non-robust features. Nevertheless, this is not always the case for debiased algorithms. A recent work uses explainability as a debugging tool to analyze debiased models (Mendelson and Belinkov, 2021). The analysis indicates that the debiased models actually encode more biases in their inner representations. It is speculated that the improved performance on the OOD data comes from the refined classification head. More research is needed to investigate whether the debiased model has captured more robust features and what is the source of their improved generalization (Rosenfeld et al., 2022).

6.5 More Challenging Evaluation Datasets

It is encouraging to see that some benchmark datasets have emerged to evaluate adversarial and OOD robustness. For example, adversarial

GLUE is proposed for adversarial robustness evaluation, which contains 14 adversarial attack methods (Wang et al., 2021a). Checklist (Ribeiro et al., 2020) and Robustness Gym (Goel et al., 2021) can be used to evaluate the robustness of LLMs. Despite these current advances, it is necessary to further curate challenging evaluation datasets: 1) covering a wider range of NLU tasks, such as reading comprehension, and 2) covering a wider range of biases, such as those listed in Section 2.1.

7 Connections to Other Directions

In this section, we provide a further discussion of the connection of shortcut learning and robust learning with other closely relevant directions.

7.1 Domain Adaptation & Generalization

The robust learning approaches that we have discussed in Section 5 are closely relevant to domain adaptation and domain generalization. The three directions share the similarity that the training and test sets are not from the same distribution, i.e., there is a certain distribution shift. However, the objective of robust learning is distinct from domain adaptation, which aims to generalize to a specific target domain (Teney et al., 2020). In contrast, robust learning is closer to domain generalization, where both areas have the goal of generalizing over a range of unknown conditions (Wang et al., 2021b). It is also worth noting that various types of dataset distribution shift could cause domain generalization problem (Wiles et al., 2022), where spurious correlation is only one of them.

7.2 Long-Tailed Classification

Long-tailed classification addresses the problem of long-tailed distributed data, where the head class contains abundant training samples and the tail class has only a few training samples (Li et al., 2022a). Shortcut learning can be treated as a special case of long-tailed classification, where easy samples correspond to the head class and hard samples represent the tail class. Some of the robust learning solutions (e.g., reweighting) in Section 5 share a philosophy similar to that of approaches to the long-tailed classification problem. Leveraging ideas from approaches to long-tailed classification could further boost the robustness of LLMs.

7.3 Algorithmic Discrimination

Shortcut learning could also lead to discrimination and unfairness in deep learning models (Du et al., 2020). In contrast to the general bias captured by the models, the spurious patterns here usually correspond to societal biases in terms of humans (e.g., racial bias and gender bias). Here, the models have associated the fairness-sensitive attributes (e.g., ZIP code and surname) with main prediction task labels (e.g., mortgage loan rejection). At the inference time, the model would amplify the bias and show discrimination towards certain demographic groups, e.g., African Americans and females.

7.4 Backdoor Attack

The previous sections focus on discussing the setting in which LLMs have unintentionally captured undesirable shortcuts. However, the adversary can intentionally insert shortcuts into LLMs, which could be a potential security threat to the deployed LLMs (Yu et al., 2021). This is termed the backdoor attack (or poisoning/Trojan attack). Backdoor attackers insert human-crafted easy patterns that serve as shortcuts during the model training process, explicitly encouraging the model to learn shortcuts (Kurita et al., 2020). Representative examples include modifying the style of text (Qi et al., 2021), adding shortcut unigrams such as double quotation marks (Du et al., 2021a), etc.

7.5 Watermarking

Different from the malicious usage of shortcut learning as the backdoor attack, shortcut learning can also be used for benign purposes. In particular, trigger patterns can be inserted as watermarks by model owners during the training phase to protect the IP of companies (Tang et al., 2020). When LLMs are used by unauthorized users, shortcuts in the format of trigger patterns can be used to claim ownership of the models.

8 Conclusions

In this article, we provide a comprehensive analysis of the LLM's shortcut learning issue for NLU tasks. Our analysis indicates that shortcut learning can be attributed to the skewed dataset, the model architecture, and the model learning dynamics. We further summarize the mitigation solutions that can be used to alleviate shortcut learning and increase the robustness of LLMs. Furthermore, we provide discussions of directions that deserve further effort from the research community and also point out the connections of shortcut learning and robust learning with other relevant directions.

Limitations

In this work, we focus primarily on LLMs with the pre-training and fine-tuning paradigm. We cover few literature that address the robustness issue for other models and training paradigms, including the fully supervised paradigm with shallow models such as LSTM and prompt-based learning with language models (Liu et al., 2021b). Additionally, we focus primarily on a few NLU tasks, such as NLI, QA, and reading comprehension. In contrast, many other types of NLP tasks are not covered in this work, such as machine translation. The reason is that these NLU tasks typically require high-level semantic understanding and reasoning, and thus suffer the most from the shortcut learning issue.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*.

Ruben Branco, António Branco, João Silva, and João Rodrigues. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Sébastien Bubeck and Mark Sellke. 2021. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems (NeurIPS)*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *Empirical Methods in Natural Language Processing (EMNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Chrysanne DiMarco and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021a. Towards interpreting and mitigating shortcut learning behavior of nlu models. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. 2021b. What do compressed large language models forget? robustness challenges in model compression. *arXiv preprint arXiv:2110.08419*.

Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *NAACL demo*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Donghoon Han, Juho Kim, and Alice Oh. 2020a. Reducing annotation artifacts in crowdsourcing datasets for natural language processing. In *The eighth AAAI Conference on Human Computation and Crowdsourcing*. AAAI.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. *Findings of EMNLP*.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020b. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *2019 EMNLP workshop*.

Benjamin Heinzerling. 2019. Nlp's clever hans moment has arrived. *The Gradient*.

Katherine Hermann and Andrew Lampinen. 2020. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems (NeurIPS)*.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. *International Conference on Learning Representations (ICLR)*.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pre-trained models. *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. 2021. Why machine reading comprehension models learn shortcuts? *ACL Findings*.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*.

Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. 2022a. Trustworthy long-tailed classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tianda Li, Ahmad Rashid, Aref Jafari, Pranav Sharma, Ali Ghodsi, and Mehdi Rezagholizadeh. 2021. How to select one among all? an empirical study towards the robustness of knowledge distillation in natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Zhiming Li, Yanzhou Li, Tianlin Li, Mengnan Du, Bozhi Wu, Yushi Cao, Xiaofei Xie, Yi Li, and Yang Liu. 2022b. Unveiling project-specific bias in neural code models. *arXiv preprint arXiv:2201.07381*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021a. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. *57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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*.

Ana Marasović. 2018. Nlp's generalization problem, and how researchers are tackling it. *The Gradient*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Debiasing classifier from biased classifier. *Advances in Neural Information Processing Systems (NeurIPS)*.

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 (ACL)*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *57th Annual Meeting of the Association for Computational Linguistics (ACL).*

Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron C Wallace. 2021. Combining feature and instance attribution to detect artifacts. *arXiv preprint arXiv:2107.00323.*

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180.*

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. 2021. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP.*

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR).*

Ahmad Rashid, Vasileios Lioutas, and Mehdi Rezagholizadeh. 2021. Mate-kd: Masked adversarial text, a companion to knowledge distillation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL).*

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *58th Annual Meeting of the Association for Computational Linguistics (ACL).*

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in Neural Information Processing Systems (NeurIPS).*

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2022. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856.*

Walid Saba. 2021. Machine learning won't solve natural language understanding. *The Gradient.*

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. *International Conference on Learning Representations (ICLR).*

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *Empirical Methods in Natural Language Processing (EMNLP).*

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems (NeurIPS).*

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. 2022. Gradient matching for domain generalization. *International Conference on Learning Representations (ICLR).*

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391.*

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.*

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *Empirical Methods in Natural Language Processing (EMNLP).*

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. *AAAI Conference on Artificial Intelligence (AAAI).*

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training. *Empirical Methods in Natural Language Processing (EMNLP).*

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *International Conference on Machine Learning (ICML).*

Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. 2020. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proceedings of the 26th ACM SIGKDD International*

Conference on Knowledge Discovery & Data Mining (KDD).

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2020. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894.*

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *North American Chapter of the Association for Computational Linguistics (NAACL).*

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 (TACL).*

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. *58th Annual Meeting of the Association for Computational Linguistics (ACL).*

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing nlu models from unknown biases. *Empirical Methods in Natural Language Processing (EMNLP).*

Prasetya Ajie Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. *Empirical Methods in Natural Language Processing (EMNLP).*

Gal Vardi, Gilad Yehudai, and Ohad Shamir. 2022. Gradient methods provably converge to non-robust networks. *arXiv preprint arXiv:2202.04347.*

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *International Conference on Learning Representations (ICLR).*

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021a. Adversarial glue: A multitask benchmark for robustness evaluation of language models. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).*

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. 2021b. Generalizing to unseen domains: A survey on domain generalization. *International Joint Conference on Artificial Intelligence (IJCAI).*

Tianlu Wang, Diyi Yang, and Xuezhi Wang. 2021c. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736.*

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. 2022. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations (ICLR).*

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *North American Chapter of the Association for Computational Linguistics (NAACL).*

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL).*

Yadollah Yaghoobzadeh, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. 2019. Robust natural language inference models with example forgetting. *arXiv e-prints*, pages arXiv–1911.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. *60th Annual Meeting of the Association for Computational Linguistics (ACL).*

Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021. Indiscriminate poisoning attacks are shortcuts. *arXiv preprint arXiv:2111.00898.*

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. *58th Annual Meeting of the Association for Computational Linguistics (ACL).*

A Further Discussions

In this section, we provide a further discussion of the robustness challenges of LLMs.

A.1 Other Training Paradigms

In this survey, we focus on characterizing the shortcut learning problem of the pre-training and fine-tuning training paradigm. Other training paradigms also suffer from the shortcut learning problem. For example, recent studies indicate that the few-shot prompt-based training paradigm also suffers from the shortcut learning problem (Utama et al., 2021). The main reason is that most existing training paradigms belong to the general data-driven training paradigm, which naturally tends to rely on dataset artifacts as shortcuts for prediction.

A.2 Other NLU-relevant Tasks

Beyond NLU tasks, other NLU-relevant tasks in real-world applications also suffer from the low robustness issue (Li et al., 2022b). Take the visual commonsense reasoning task for example, the models exploit the co-occurring text between input (question) and output (answer options) to make predictions, rather than performing the desired visual reasoning task (Ye and Kovashka, 2021). This has also been observed in other vision language tasks, such as VQA (Niu et al., 2021). The similarity for these tasks is that successful predictions rely on human-like reasoning and sometimes might also rely on world knowledge. Therefore, we need to carefully interpret the results, especially when it is claimed that models with pure data-driven training outperform the performance of humans.

A.3 IID and Robustness Trade-off?

Another open question is about the connection between IID performance and OOD robustness performance. To the best of our knowledge, there are no consistent observations. For example, there is a linear correlation between IID performance and OOD generalization for different types of models introduced in Section 4.2. On the contrary, most robust learning methods introduced in Section 5 will sacrifice IID performance, although some of them could preserve IID performance. It deserves further research on the conditions under which the trade-off would occur. These insights could help the research community design robust learning frameworks that can simultaneously improve OOD and IID performance.