A Survey on Out-of-Distribution Evaluation of Neural NLP Models

Xinzhe Li¹, Ming Liu¹, Shang Gao¹ and Wray Buntine²

¹Deakin University
²VinUniversity

{lixinzhe, m.liu, shang.gao}@deakin.edu.au, wray.b@vinuni.edu.vn

Abstract

Adversarial robustness, domain generalization and dataset biases are three active lines of research contributing to out-of-distribution (OOD) evaluation on neural NLP models. However, a comprehensive, integrated discussion of the three research lines is still lacking in the literature. In this survey, we 1) compare the three lines of research under a unifying definition; 2) summarize the data-generating processes and evaluation protocols for each line of research; and 3) emphasize the challenges and opportunities for future work.

1 Introduction

Neural NLP models have been reported for their superhuman performance on many language understanding and generation tasks, such as sentiment analysis and machine reading comprehension (MRC). Recent studies show that these models lack human-level understanding of language since they are vulnerable beyond in-distribution (ID) test data and fail to generalize on out-of-distribution (OOD) data, such as perturbed examples under adversarial attacks [Ebrahimi et al., 2018], text from different domains [Hendrycks et al., 2020] and examples against dataset biases [Gururangan et al., 2018], which lie in the three mainstream lines of research: adversarial robustness, domain generalization and dataset biases.

Despite recent progress in each of the three research lines, there has not been a survey that comprehensively reviews and investigates the connections between these lines. We have the following motivations for such an investigation: 1) explaining some phenomena across the three lines of research. For example, methods designed to improve model generalization on one OOD type can defend against other types of OOD data [Yi et al., 2021] but sometimes degrade model performance on other OOD types [Bras et al., 2020; Gokhale et al., 2022]; 2) encouraging future researchers to perform a comprehensive OOD evaluation while proposing their new methods; 3) suggesting that all the OOD types be addressed thoroughly in future research.

In this survey, we first unify the three research lines as the study of distribution shift, which gives us a starting point to find connections between them in terms of shifted features (§2). Furthermore, we summarize their differences in OOD data generation and evaluation protocols (§3). In particular, we categorize different methods to generate OOD data, including data with natural domain shift (NDS) for domain generalization, debiased data for evaluating dataset biases and adversarial examples for adversarial robustness. We then categorize evaluation protocols for the OOD data into two classes: data-based and method-based evaluation. Finally, we identify opportunities based on the discussion of connections and differences between the three OOD types, including developing a comprehensive benchmark for OOD evaluation, advising caution about breaking the assumptions of covariate shift, improving general OOD performance and using the detection methods across different OOD types. Also, we demonstrate the gap in adversarial robustness for future work (§4).

2 Definition

This section first introduces a comprehensive definition of distribution shift across the three research lines and then shows how the definition aids in elucidating the interconnections regarding the shifted features.

2.1 Distribution Shift

Deep Neural Network (DNN) $f(\theta)$ can achieve remarkable or even superhuman performance on NLP benchmarks, such as GLUE with in-distribution (ID) test data $D_{test} \sim P_0$, when $f(\theta)$ is trained on data sampled from the data-generating distribution (an unknown distribution under a data-generating process) $D_{train} \sim P_0$. However, the well-trained model often fails to generalize to various unforeseen examples. Formally, the unforeseen data (or OOD data) $D$ can be characterized by a shifted distribution $\tilde{P}$. The distribution shift includes covariate shifts (input distribution shifts) $P_0(\mathcal{X}) \neq \tilde{P}(\mathcal{X})$ and label distribution shifts $P_0(\mathcal{Y}) \neq \tilde{P}(\mathcal{Y})$, where $\mathcal{X}$ is the input space, and $\mathcal{Y}$ is the ground-truth output space. We unify the three lines of research as a study of covariate shifts. For brevity, the input notation $\mathcal{X}$ is sometimes omitted when referring to it, as exemplified by the use of $P_0$ to represent $P_0(\mathcal{X})$.

Domain generalization and dataset biases. Real-life documents in different domains, e.g., news and fiction text [Hendrycks et al., 2020], are typically characterized by systematically distinct properties and originate in different data-
generating distributions. Domain generalization investigates such natural domain shifts, and relevant literature often uses domain shift and distribution shift interchangeably. Some literature [Arora et al., 2021; Gokhale et al., 2022] also defines that data with particular biases and those without such biases come from different domains. Therefore, the studies of domain generalization and dataset biases can be intuitively defined in the framework of distribution shifts.

**Adversarial robustness: from robustness to distribution shift.** Initially, adversarial methods evaluate model robustness against small perturbations in the worst-case scenarios, especially for continuous data like images [Goodfellow et al., 2015]. Hence, prior studies typically separate adversarial robustness from the study of domain generalization and distribution shift. We can unify adversarial robustness as a study of distribution/domain shift. Conceptually, we can define the shifted distribution in the adversarial setting as a distribution around the original distribution \( P_0 \). We perform a perturbation on text \( x \in P_0(\mathcal{X}) \) to simulate OOD data from the shifted distribution. In other words, adversarial perturbations encode subtle domain shifts, as discussed in §2.3, and the shifted domains are characterized by the adversarial methods summarized in §3.3.

### 2.2 Shifted Features

To concretize the shifted features, we first define task-relevant features \( s_{rel} \) and task-irrelevant features \( s_{ir} \), which satisfy \( P(s_{rel}|y) \neq P(s_{ir}) \) and \( P(s_{ir}|y) = P(s_{ir}) \), respectively. \( P \) is an arbitrary distribution.

**Background features and semantic features.** When \( P \) is the ground-truth distribution \( P_{true} \), Arora et al., [2021] name \( s_{rel} \) as semantic features since most realistic NLP tasks address the semantics of text, and \( s_{ir} \) as background features, which can be the syntax, writing style or task-irrelevant text.

**Biased features are task-irrelevant features under \( P_{true} \) despite being learned as task-relevant features from \( P_0 \).** \( P_0(\mathcal{X}) \) can be factorized into the conditional distributions of generalized features \( g(\mathcal{X}) \) and biased features \( b(\mathcal{X}) \).

\[
P_0(\mathcal{X}) = \sum_y P_0(g(\mathcal{X})|y)P_0(b(\mathcal{X})|y)P_0(y),
\]

assuming conditional independence between \( b(\mathcal{X}) \) and \( g(\mathcal{X}) \) given \( y \). By definition, \( g(\mathcal{X}) \) should always be task-relevant features in the true distribution \( P_{true} \), i.e., \( g(\mathcal{X}) \in s_{rel} \), while biased features \( b(\mathcal{X}) \) are only predictive for ID dataset \( D \sim P_0 \). We can express \( b(\mathcal{X}) \) as \( \{s|s \in P_0 \text{ and } s \notin P \} \).

### 2.3 Shifted Features in Three OOD Types

This section shows how each OOD type can potentially shift the three kinds of features.

**Shifted features on adversarial examples.** Shifted features of adversarial examples depend on the attack methods. White-box and grey-box methods compute gradients or output logits, respectively, while black-box methods can only query final predictions. With the guidance of gradients, white-box methods tend to shift the biased features since the input gradients can help identify the biased features for perturbations [Ilyas et al., 2019; Wallace et al., 2019a]. Perturbation types largely affect whether it is a semantic shift or background shift (See §3.3). For example, character editing may lead to unseen words, e.g., “wonderful” to “wonderful”, which belongs to semantic shift. In contrast, black-box methods explicitly change background features at the sentence level. For example, Qi et al., [2021] paraphrase with different writing styles.

**Shifted features on debiased data.** Debiased data, by definition, is generated by removing biased features. However, the generation process often encodes either background features \( s_{ir} \) or semantic features \( s_{rel} \). For example, debiased data generated by syntactic templates [McCoy et al., 2019] encode background shift, while MRC data requiring numerical reasoning to avoid dataset biases [Naik et al., 2018] change semantic features.

**Shifted features on NDS data.** We classify natural domain shift into two categories, namely genres and sub-populations. Both background and semantic shifts can occur in both categories, such as different vocabularies associated with sub-populations or genres [Barrett et al., 2019]. For example, for sentiment analysis, Tweets tend to express sentiments with emotions compared to the genre of movie reviews, and sub-populations can differ in the way to express emotions [Oprea and Magdy, 2020]. Furthermore, DNNs can easily learn background features for each sub-population or genre as biased features. Writing styles can also serve as another background feature that changes across genres [Arora et al., 2021], e.g., text from fiction books to Wikipedia content.

### 3 OOD Performance Evaluation

This section compares the concrete methodologies of OOD data generation and evaluation protocols for each line of research. Their differences are summarized in Table 1.

#### 3.1 NDS Data Generation

NDS data can be generated from different genres or sub-populations. The former can be acquired from various data sources, whereas the latter involves partitioning data into sub-populations based on attributes of entities or individuals. Diverse sets of data genres may display a fusion of unique textual styles, syntactic structures, and lexicons.

**Genres.** Different genres refer to text written by different authors or annotators or for different audiences, or come from different data sources. For example, academic papers are written in formal language while online reviews on Yelp contain non-standard orthography. Since a training dataset typically comes from a specific data source and contains only one genre, even well-trained models generalize poorly to different genres. Although some datasets are created to include multiple genres of text, models still perform worse beyond the coverage of these genres. For example, MultiNLI for NLI consists of text from ten distinct genres. However, even the large pretrained language models cannot generalize well to a different genre [Arora et al., 2021], such as SNLI from image captions or WNLI from fiction books. Take text classification as another example. Classifiers cannot perform well on
text from disjoint annotators, e.g., SST data from experts v.s. lengthy IMDB reviews from laymen [Hendrycks et al., 2020; Arora et al., 2021]. Also, MRC models trained on Wikipedia hardly generalize on data from New York Times articles, Reddit posts, and Amazon product reviews.

Sub-populations. Multiple studies identify NDS data by dividing data into sub-populations according to different attributes of objects or individuals obtained from metadata. Arora et al., [2021] select reviews from different businesses, e.g., restaurant reviews v.s. movie reviews. For the same business, Hendrycks et al., [2020] split them according to different types of products (e.g., women’s clothing, shoes) or restaurants (e.g., American v.s. Chinese restaurants). Furthermore, model performance can vary on data with different demographics, which is closely connected to another line of research about fairness. For example, Borkan et al., [2019] find the worst test accuracy on non-toxic comments from the black population.

3.2 Debiased Data Generation
In this section, we first divide dataset biases into two categories and introduce a text summarization dataset bias that does not fall under either one. We then classify data-generating methods into two categories according to the knowledge of dataset biases. Table 2 and 3 summarize dataset biases and debiased datasets for various NLP tasks, respectively.

### Dataset Biases

**Lexical correlations.** Words frequently appearing in the training examples of a particular class can be predictive biases. For example, the word “American” might bias a model to perform well on American restaurants but poorly on Chinese restaurants. Similarly, “women” might bias a model to perform well on women’s clothing but poorly on men’s clothing.

**Lexical overlap.** Words that are frequently used in both the premise and hypothesis can be predictive biases. For example, the word “restaurant” might bias a model to perform well on restaurant reviews but poorly on movie reviews.

**Positional bias.** Words that are frequently used in certain positions can be predictive biases. For example, the word “American” might bias a model to perform well on American restaurants but poorly on Chinese restaurants.

### Methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Tasks</th>
<th>Debiased Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biased inspired</td>
<td>TC</td>
<td>c-IMDB [Kaushik et al., 2020]</td>
</tr>
<tr>
<td></td>
<td>NLI</td>
<td>HANS [McCoy et al., 2019], Stress Text [Naik et al., 2018]</td>
</tr>
<tr>
<td></td>
<td>MRC</td>
<td>Para-SQUAD [Lai et al., 2021], Adv-SQUAD [Jia and Liang, 2017]</td>
</tr>
<tr>
<td></td>
<td>Paraphrase</td>
<td>PAWS [Zhang et al., 2019]</td>
</tr>
<tr>
<td>Systematic</td>
<td>Fact Checking</td>
<td>FEVER-Symmetric [Schuster et al., 2019]</td>
</tr>
<tr>
<td></td>
<td>NLI</td>
<td>SNLI-APLITE [Bras et al., 2020]</td>
</tr>
<tr>
<td></td>
<td>Reasoning</td>
<td>Winogrande [Sakaguchi et al., 2020]</td>
</tr>
<tr>
<td></td>
<td>CGI</td>
<td>Swag [Zellers et al., 2018]</td>
</tr>
</tbody>
</table>

Table 2: Debiased datasets for various NLP tasks. TC: Text Classification; NLI: Natural Language Inference; MRC: Machine Reading Comprehension; Paraphrase: Paraphrase Identification; Reasoning: Commonsense Reasoning; CGI: Commonsense Grounded Inference.
ases for that class. For example, when NLI annotators create a majority of contradictory hypotheses by negating premises, models learn the spurious correlation between negation words (‘no’, ‘never’) and labels (hypothesis-only reliance [Gururangan et al., 2018]). The success of class-wise adversarial attacks comes from exploring these predictive words to some extent [Wallace et al., 2019a]. Similarly, claim-only models can perform well on ID test data without the context of evidence for fact-checking due to predictive words [Schuster et al., 2019]. MRC models can find the answer in a given paragraph by matching the question type, e.g., “November 2014” for a “when” question (question word matching [Lai et al., 2021] or entity type matching [Sugawara et al., 2018]), and hence they are easily distracted to the wrong answers on OOD data [Jia and Liang, 2017]. Emelin et al., [2020] observe that the overlap between the premise and hypothesis leads to the “en-cator for sentence-pair classification. For example, the high lexical overlap. Lexical overlap is another predictive indica-tor for text-pair classification. For natural language inference (NLI), McCoy et al., [2019] specify linguistic phenomena (constituent and subsequent heuristics) behind lexical overlap. They devise heuristics-based syntactic templates to create sentence pairs that have high lexical overlap but contradict each other. Kaushik et al., [2020] construct counterfactual examples by breaking spurious correlations in sentiment analysis datasets (i.e., IMDB) and NLI datasets. They annotate the text into the targeted label by largely keeping the original text. In this way, the dataset biases of the original label are kept. To break the lexical overlap between paraphrase pairs, Zhang et al., [2019] create non-paraphrase sentences via word scrambling. This method changes the meanings of paraphrases but keeps the overlapping words. For MRC, Jia and Liang [2017] add a sentence containing words overlap-ping with the question into paragraphs and find that models tend to select answers from the sentence. They also combine the bias-inspired approach with a grey-box attack to generate adversarial examples. Gardner et al., [2020] rely entirely on experts who have the knowledge of dataset biases to generate debiased data.

**Systematic approaches.** Debased data can also be generated by bias mitigation techniques without knowing dataset biases. For example, Bras et al., [2020] use a simple classifier to identify biased examples and generate debiased datasets for NLI, commonsense reasoning and grounded commonsense inference, respectively. Although some works try to train robust models with such filtered data, filtering always leads to significant drops on ID data and even OOD data [Gokhale et al., 2022]. Therefore, it can be more practical to use them for OOD evaluation.

### 3.3 Adversarial Example Generation

Given a well-trained model, adversarial examples are generated by modifying the given text (or reference data) to make the model output wrong predictions. Typically, the generation process satisfies semantics-preserving and label-preserving assumptions. Mostly, semantics preservation can guarantee the same label for a perturbed text. Thus, a robust model should generate invariant outputs $\arg \max_{y^*} f(x) = y^*$ where $y$ is the ground-truth label and $\tilde{x}$ is the perturbed text. Adversarial attacks have been widely studied to automatically generate adversarial examples. We categorize the attack methods according to their perturbation types, perturbation space and adversary’s knowledge. Table 4 exemplifies typical methods for each category.

### Perturbation Types

There are four common perturbation types that exhibit different granularities, namely, character editing, word substitution, paraphrasing and universal adversarial perturbation (UAP).

### Character editing mimics real-life accidental typos or spelling variants in social media via character swapping (e.g., “place” ⇒ “plcae”, deletion (e.g., “artist” ⇒ “arlist”), insertion (e.g., “computer” ⇒ “computor”) and substitution (e.g., “computer” ⇒ “computor”) [Gao et al., 2018]. Humans are robust to spelling errors maintaining certain morphological or phonological characteristics, e.g., “computer”. Hence, a few edits on a word would not affect the human perception of the word but may lead to completely opposite predictions by the models. In contrast, word substitutions require semantics-preserving constraints on substitute words or rely on a vocabulary of synonyms for substitutions [Zang et al., 2020].

### Paraphrasing is seldom considered in adversarial attacks due to the difficulty of generating high-precision paraphrases.

### The last type of adversaries follows the work to generate UAPs for images, which perturb any images of a particular class [Moosavi-Dezfooli et al., 2017]. Behjati et al., [2019] generate the UAPs for NLP models in the form of n-grams (i.e., non-sensical phrases), which lead to misclassification when appended to text of a class. Wallace et al., [2019a;
<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Perturbation Space</th>
<th>Perturbation Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>White-box</td>
<td>Continuous</td>
<td>Word substitution [Gong et al., 2018], UAP [Song et al., 2021], Character editing [Liu et al., 2022]</td>
</tr>
<tr>
<td></td>
<td>Discrete</td>
<td>Word substitution [Li et al., 2018], Ebrahimi et al., 2018], UAP [Wallace et al., 2019a; Behjati et al., 2019]</td>
</tr>
<tr>
<td>Grey-box</td>
<td>Discrete</td>
<td>Word substitution [Alzantot et al., 2018; Zang et al., 2020; Jin et al., 2020; Li et al., 2020], Character editing [Gao et al., 2018]</td>
</tr>
<tr>
<td>Black-box</td>
<td>Continuous (Only generation model)</td>
<td>Paraphrasing [Iyyer et al., 2018; Ribeiro et al., 2018; Qi et al., 2021]</td>
</tr>
</tbody>
</table>

Table 4: Classification of Attack Methods Based on Adversaries’ Knowledge, Perturbation Space, and Perturbation Types. UAP refers to Universal Adversarial Perturbation (UAP). Several studies employ generation models to produce paraphrases without adding any noise for perturbations.

2020] even extend the idea of UAPs for MRC, language modeling and machine translation. These studies assume that inserting non-sensical phrases does not change the semantics of text and ground-truth labels. Song et al., [2021] generate natural phrases (e.g., “natural energy efficiency”) as UAPs via Adversarially Regularized Auto Encoder (ARAE). However, it can only maintain the coherence of UAPs rather than the whole perturbed text.

Perturbation Space
A text consists of a sequence of tokens (e.g., words or subwords), denoted as \( x = w_1, w_2, \ldots, w_T \), where \( T \) is the number of tokens in the vocabulary \( V = \{ w_1, w_2, \ldots, w_K \} \) (\( K \) is the size of the vocabulary). Note that the vocabulary can differ from the ones of the victim models, especially under black-box attacks.

All types of perturbations, except for paraphrasing, can be perturbed in a discrete space. When perturbing via word substitutions or UAPs, the objective is to identify significant tokens in \( x \) and substitute them with tokens from the vocabulary \( V \) in order to generate incorrect model predictions. This is a combinatorial optimization problem, where the size of the search space is \( K^T \). The search space is intractable and is typically addressed using heuristic-based approaches [Alzantot et al., 2018; Zang et al., 2020] or approximation methods [Ebrahimi et al., 2018; Wallace et al., 2019a]. Character editing also requires identifying significant tokens, but it involves a subsequent manipulation of characters within the identified tokens [Gao et al., 2018]. Assuming the average token length is \( l \), the search space becomes \( l^T \), leading to combinatorial explosion of modifications.

There are two types of methods to perturb text in the continuous space: 1) perturbing tokens (characters/subwords/words) in the embedding space. Gong et al., [2018] add continuous perturbations directly to the continuous representation \( e_w \) of token \( w \) via element-wise addition between the perturbation \( \eta \) and \( e_w \), similar to pixel-wise addition for images [Goodfellow et al., 2015]. To transform \( \eta + e_w \) back to the text, they search for the nearest token \( \tilde{w} \) in the embedding space where \( e_{\tilde{w}} + \eta \in V \). However, it probably generates a semantically variant token, because the closest and most legible token can be far away from the original one in projection space. Note that they still require searching for important tokens. 2) reparameterization for paraphrasing or UAPs. Zhao et al., [2018] reparameterize the text \( x \) into \( z \) in a continuous space and then perturb \( z \) into \( \tilde{z} \). They also train a text generator (e.g., LSTM) to decode \( \tilde{z} \) back to text. Song et al., [2020] apply the reparameterization trick for UAP generation.

Moreover, generation models are commonly used to create paraphrases such as machine translation models for back-translation [Ribeiro et al., 2018], a syntactically controlled paraphrase network [Iyyer et al., 2018] or a text style transfer model [Qi et al., 2021]. The latter two specifically change task-irrelevant features.

Adversary’s Knowledge
We can categorize the attack methods into three types (white-box, grey-box and black-box) according to the three levels of model knowledge: model parameters, output logits (i.e., estimated probability distribution) and final predictions.

The white-box methods require model parameters to propagate gradients back to the input (e.g., the gradient of an adversarial loss with respect to the input word embedding \( \nabla_{w}\mathcal{L} \)). They need \( \nabla_{w}\mathcal{L} \) to find important tokens and generate perturbations, e.g., the noise in the embedding space or substitute tokens. Li et al., [2018] adapt the gradient magnitude to find important words for perturbations and search for substitute words for each targeted label since the magnitude in each dimension of the word embedding indicates the sensitiveness of the prediction to the change. Gong et al., [2018] directly use gradients to perform perturbations on the embedding space. Ebrahimi et al., [2018] and Wallace et al., [2019a; 2020] use the gradients and word embeddings to approximate the loss for word substitutions. Specifically, they approximate the loss change of substituting a word \( w \) with another word \( s \) in the vocabulary by the inner product of the word embedding \( e_w \) and \( \nabla_{e_w} \mathcal{L} \), where \( \mathcal{L} \) is the adversarial loss. The selected word \( s \) is expected to minimize the adversarial loss on the perturbed text \( \tilde{x} \).

\[
\arg\min_s \mathcal{L}(\tilde{x}, y) \approx e_s^T \nabla_{e_w} \mathcal{L}(x, y) \tag{2}
\]

Under the grey-box setting, an adversary only has access to the probability distribution (or logits) over the output vari-
able \( Pr(y|x) = f(x) \), where \( Pr(y|x) \) is the probability of the outcome \( y \) given the text \( x \). Grey-box methods employ the probabilistic information to measure how important each token \( w_i \in x \) is to the prediction. For \( w_i \), its importance score \( I_i \) is commonly calculated as the difference of the probabilities of the correct class \( Pr(y_{true}|x) - Pr(y_{true}|x_{\backslash w_i}) \) \cite{Li et al., 2020}, where \( x_{\backslash w_i} \) can either be \( \{ w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_T \} \) (delete \( w_i \)) \cite{Li et al., 2019; Jin et al., 2020} or \( \{ w_1, \ldots, UNK, \ldots, w_T \} \) (replace \( w_i \) with the unknown token \( UNK \)) \cite{Gao et al., 2018; Li et al., 2020}. Jin et al., [2020] also add the change to the misclassified class \( f_\hat{y}(\hat{x}) - f_y(x) \) if \( \arg\max_{\hat{y}} f(\hat{x}) \neq \arg\max f(x) \), where \( \hat{y} \) is the misclassified class, and \( f_\hat{y}(\hat{x}) \) is the probability of the misclassified class for the perturbed sample. Afterward, grey-box methods apply heuristics to search for substitute words or evaluate noisy words from character editing via \( Pr(y_{true}|x) \).

The black-box attacks require no information about the victim model and commonly generate paraphrases. We have discussed the concrete methods in the previous paragraphs for perturbation types and space.

### 3.4 Evaluation Protocols

**Data-based evaluation.** There are different evaluation protocols for the three data types. NDS data and debiased data are evaluated on the same aggregate metrics for ID data (e.g., accuracy for text classification and F1 scores for MRC). The drops of these metrics from ID data to OOD data reveal the model’s ability of OOD generalization.

**Method-based evaluation.** It is meaningless to apply the ID metrics to adversarial examples unless used for adversarial transferability, since adversarial robustness only concerns the misclassified examples for the victim model. Instead, adversarial robustness is evaluated by the effectiveness of an attack method generating adversarial examples. The common evaluation metrics are attack success rate and the number of queries. Specifically, given a test sample \( D_{test} \), attack success rate measures the proportion of the data points in \( D_{test} \) which can be perturbed to fool a victim model. It is calculated by dividing the number of adversarial examples that successfully achieve attack goals, e.g., misclassification, by the size of \( D_{test} \). The number of model queries during an attack measures how many perturbed examples have been evaluated before a valid adversarial example appears. Besides the query for predictions, white-box and grey-box methods also query victim models to get gradients or output logits. These two metrics can be used to compare the adversarial robustness between models (with fixed attack methods) \cite{Gokhale et al., 2022} and the performance of different attack methods \cite{Li et al., 2020} (with fixed victim models).

**Adversarial transferability.** Adversarial transferability is another evaluation type for adversarial robustness but uses the data-based evaluation method. It generates adversarial examples from a base model and then uses them to evaluate other models. The vulnerability can transfer across different model architectures and datasets \cite{Wallace et al., 2020}. Liang et al., \cite{2021} show that transferability can even occur across tasks as long as the source data distribution and the target one are close. Adversarial transferability encourages the generation of costly human-made adversarial examples since they can be helpful in evaluating and improving the robustness of various models. For example, Nie et al., \cite{2020} and Wallace et al., \cite{2019b} show model predictions or interpretations (e.g., highlighting words that are important for model predictions) to crowdworkers to assist in the generation of adversarial examples.

### 4 Opportunities and Challenges

This section provides insights for future work based on the above OOD discussions. The opportunities and challenges related to adversarial robustness are also investigated.

#### 4.1 General Challenges and Opportunities

Our study provides insights into OOD evaluation, OOD generalization and OOD detection.

**A benchmark for all the OOD types.** Developing a comprehensive benchmark for all the OOD types can help evaluate the model’s actual ability of language understanding and identify potential incapacity. Although there are open-source tools developed for OOD evaluation, most of them can only evaluate one perspective of the distribution shift. It is challenging to develop such a benchmark due to the idiomatic lock-in in each line of research, different data-generating processes and evaluation protocols. The disclosure of their connections and differences in this survey can provide insights into developing such a comprehensive benchmark for OOD evaluation. For example, adversarial robustness can be evaluated via adversarial transferability to unify evaluation metrics.

**Data-generating processes may break the assumptions of covariate shift.** Although all the three research lines aim to evaluate data with covariate shifts, some processes may break the assumptions of covariate shifts, i.e., \( \mathbb{P}_0(X|Y) \) and \( \mathbb{P}_0(Y) \). These two assumptions also indicate \( \mathbb{P}_0(X|Y) \) as a special case of covariate shift due to \( \mathbb{P}_0(X|Y) = \frac{\mathbb{P}_0(X|Y)\mathbb{P}_0(Y)}{\mathbb{P}_0(Y)} \). For example, although \( \mathbb{P}_0(X|Y) \) is always assumed the same for most NLP tasks since humans usually annotate each sentence with the same label, gradient-based methods may generate adversarial examples with labels mismatching with humans’ perceptions. In addition, some methods for generating debiased data may change \( \mathbb{P}_0(Y) \). For example, Zhang et al., \cite{2019} balance examples with high lexical overlap for all kinds of labels in their debiased dataset to prevent the feature of lexical overlap from being a predictive indicator for paraphrase identification. Future work that evaluates covariate shifts (i.e., the changes of input domains) may better consider the two assumptions for data-generating processes and analyze their effects.

**Enhance general OOD performance.** The three types of out-of-distribution (OOD) data can be created by altering task-irrelevant features, as explained in Section 2. Our first argument suggests that OOD generalization techniques should aim to improve the acquisition of general linguistic or semantic knowledge to perform well on all three types of
OOD evaluation. This accounts for the observation that unsupervised pretraining and data augmentation strategies for learning general linguistic or semantic knowledge can enhance generalization on different domains and adversarial robustness [Gururangan et al., 2020; Gokhale et al., 2022]. Notably, such augmentation techniques [Wei and Zou, 2019] may solely modify the background features of training samples without any adversarial process. Our second argument is that even though dataset filtering for bias mitigation may improve model performance on biased datasets, it may lead to inferior performance on ODS data and decreased adversarial robustness, as biased data still contains knowledge of semantic and background features that can enhance general OOD performance. Supporting evidence for this argument is provided by Bras et al., [2020] and Gokhale et al., [2022] across various NLP tasks.

Utilizing detection methods across different research lines. The connections between OOD types regarding the shifted features also motivate future researchers to apply OOD detection methods across the three research lines. For example, since density estimation methods can effectively detect NDS data with background shift [Arora et al., 2021], they may defend against black-box attacks, which always generate adversarial examples by shifting background features.

4.2 Challenges and Opportunities for Adversarial Robustness

Beyond semantics preservation. Adversarial examples can be generated beyond the assumption of semantics preservation. Chen et al., [2022] define obstinate adversarial examples, which satisfy two conditions: 1) their ground-truth labels after the perturbation are changed, but 2) the victim model maintains the original predictions. Similarly, Kaushik et al., [2020] minimally perturb test examples to change the labels of test examples. There are many reasons to explore obstinate adversarial examples. Gardner et al., [2020] argue that such examples characterize the correct decision boundary for the task. Chen et al., [2022] find that models with adversarial training on these examples reveal vulnerability to obstinate adversarial examples. Also, breaking the assumption of semantics preservation allows more adversarial behaviours for tasks beyond single-sentence classification. For example, Song et al., [2020] generate nonsensical or natural sentences, leading to invariant model outputs for sentence-pair classification. Wallace et al., [2020] change malicious content into nonsensical text which still makes machine translation models translate it into bad language.

Developing realistically harmful adversarial behaviours. There are different model behaviours designed for malicious intents during adversarial attacks. For classification tasks, we can either specify the expected prediction (targeted attack) or accept any prediction different from the correct label (non-targeted attack). The harm of attack depends on the specific task or labels. For example, it is harmful to generate adversarial examples to evade the detection of misinformation or toxic content. In contrast, there are many possibilities of adversarial behaviours for generation tasks due to various combinations of input and output text. Cheng et al., [2020] define an attack where all the words in the output text are different from the original output sequence while the input is similar. This kind of attack can lower the standard evaluation metrics based on n-grams, like BLEU scores. Wallace et al., [2020] develop UAPs to make MT models hardly generate any translation or output random words. Targeted keyword attacks [Cheng et al., 2020] can make models generate targeted words. Future researchers can define more realistically harmful behaviours, such as the generation of malicious nonsense or racial materials.

Adversarial robustness may not be a good proxy for realistic scenarios. Adversarial perturbations may make text deviate from their real-world distribution and generate rarely occurred examples. Particularly, pure gradient-based methods only pursue worst-case perturbations without considering the naturalness of adversarial examples. Hence, to alleviate this problem, adversarial candidates are commonly validated by some metrics such as language model perplexity [Alzantot et al., 2018], and part-of-speech matching [Ebrahimi et al., 2018]. However, attack processes become computationally expensive by including these separate modules and rejecting most of the perturbed examples. Another type of approach is to use limited search space for adversarial perturbations, e.g., using synonyms for word substitutions with the compromise of worst-case performance [Zang et al., 2020]. It is worth exploring the attack methods that can inherently ensure the naturalness of perturbed text. There are some scenarios, as summarized below, where adversarial robustness against rare and unnatural text is undoubtedly critical. 1) When models’ behaviours can cause devastating outcomes, e.g., astronauts or legal services. 2) When models are deployed in ubiquitous, unforeseeable scenarios. For example, Facebook translation once made the mistake of translating the simple phrase “good morning” in Arabic into “attack it” in English. Such a mysterious translation convinced the police that the user posting this message might launch a vehicle attack, resulting in the arrest of the user. 3) When attackers have sufficient malicious intent to explore rare cases, such as fact verification and security tasks where attackers would like to evade the model’s detection to get unqualified access, e.g., propagating anti-social content or fake news on social media.

5 Conclusion

In this paper, we characterized and summarized the three mainstream lines of research on dataset biases, domain generalization and adversarial robustness. We encouraged future researchers to think comprehensively about OOD evaluation and improve the OOD generalization of NLP models on all the three types of OOD data. We also highlighted the gap between adversarial robustness and realistic OOD evaluation.

References


[Ilyas et al., 2019] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In Advances in NIPS, 2019.


